Protein Science (1996), 5:1594-1612. Cambridge University Press. Printed in the USA.

Ribosome-mediated translational pause and protein domain organization

T.A. THANARAJ[1,2] AND PATRICK ARGOS[1]

[1] European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D-69012, Heidelberg, Germany
[2] Centre for Cellular & Molecular Biology, Hyderabad, India

(RECEIVED February 12, 1996; ACCEPTED May 28, 1996)

Abstract

Because regions on the messenger ribonucleic acid differ in the rate at which they are translated by the ribosome and because proteins can fold cotranslationally on the ribosome, a question arises as to whether the kinetics of translation influence the folding events in the growing nascent polypeptide chain. Translationally slow regions were identified on mRNAs for a set of 37 multidomain proteins from *Escherichia coli* with known three-dimensional structures. The frequencies of individual codons in mRNAs of highly expressed genes from *E. coli* were taken as a measure of codon translation speed. Analysis of codon usage in slow regions showed a consistency with the experimentally determined translation rates of codons; abundant codons that are translated with faster speeds compared with their synonymous codons were found to be avoided; rare codons that are translated at an unexpectedly higher rate were also found to be avoided in slow regions. The statistical significance of the occurrence of such slow regions on mRNA spans corresponding to the oligopeptide domain termini and linking regions on the encoded proteins was assessed. The amino acid type and the solvent accessibility of the residues coded by such slow regions were also examined. The results indicated that protein domain boundaries that mark higher-order structural organization are largely coded by translationally slow regions on the RNA and are composed of such amino acids that are stickier to the ribosome channel through which the synthesized polypeptide chain emerges into the cytoplasm. The translationally slow nucleotide regions on mRNA possess the potential to form hairpin secondary structures and such structures could further slow the movement of ribosome. The results point to an intriguing correlation between protein synthesis machinery and in vivo protein folding. Examination of available mutagenic data indicated that the effects of some of the reported mutations were consistent with our hypothesis.

Keywords: codon usage; domain linkers; peptide channel; protein folding; ribosome; translation

The rate of protein translation is nonuniform along the mRNA (Varenne et al., 1984). Ribosomes seem to pause as well as stack at specific regions on mRNA (Chaney & Morris, 1979; Wolin & Walter, 1988; Kim et al., 1991; Kim & Hollingsworth, 1992). Two of the mRNA features that slow the ribosome during translation are the presence of slow codons (Varenne et al., 1984; Bonekamp et al., 1985; Wolin & Walter, 1988) and the formation of higher order nucleotide structures (Chaney & Morris, 1979; Tu et al., 1992). The resultant translational pause may temporally separate the adjacent regions on the growing nascent polypeptide. Given that proteins can fold cotranslationally with the possible involvement of the ribosome (Phillips, 1966; Baldwin, 1975; Yonath, 1992; Kudlicki et al., 1994; Wiedmann et al., 1994; Brimacombe, 1995), the temporal separation might exert a phenotypic effect on the structural organization of the encoded protein. We show that as much as 70% of the domain boundaries in a set of 37 *Escherichia coli* proteins with known tertiary structure are coded by slow regions (that encompass slow codons) on mRNA. Further, the analysis of amino acid usage in the corresponding structural regions on the encoded proteins suggests a second pause, viz. in the transit of the nascent peptide in the ribosome cavity (Yonath, 1992).

Reprint requests to: T.A. Thanaraj, Research Officer, Applied Biocomputing & Bioinformatics, European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, U.K.; e-mail: thangave@icgeb.trieste.it.

Results and discussion

Rationale

The speed of ribosome travel on mRNA may well depend on several parameters such as codon usage, codon-anticodon interactions (Grosjean & Fiers, 1982), codon context (Lipman & Wilbur, 1983; Yarus & Folley, 1985; Shpaer, 1986; Varenne et al., 1989), adjacent codons (Gutman & Hatfield, 1989), and second-

ary structure of the mRNA region. Among these, the supposition regarding the correlation between codon usage and overall translation rate is supported by various lines of published experimental evidence, as will be described subsequently. Theoretical models have been developed to explain ribosome "jam" and "queue" caused by clusters of rare codons and the overall effect on the translation efficiency (Zhang et al., 1994) when the rare codons are located at different regions on the mRNA.

It is known that cellular levels of the isoaccepting tRNA molecules for an amino acid are not identical (Ikemura, 1981) and this differential population has been suggested to control the relative translational rate of the cognate synonymous codons (Ikemura & Ozeki, 1983; Ikemura, 1985; Sorensen et al., 1989; Sorensen & Pedersen, 1991). The mRNAs of highly expressed genes, supposed to have higher rates of translation so that the encoded proteins are synthesized abundantly, preferentially use a subset of codons that maximizes the rate of translation (Ikemura & Ozeki, 1983; Ikemura, 1985; Sharp & Li, 1987; Andersson & Kurland, 1990; Lobry & Gautier, 1994). It is well documented that some codons are translated more slowly than others and abundant codons are mostly translated at a higher rate (Carter et al., 1986; Harms & Umbarger, 1987; Bonekamp et al., 1989; Curran & Yarus, 1989; Sorensen et al., 1989; Sorensen & Pedersen, 1991). It has been further suggested that there is some optimization in the choice of amino acid types used in abundant proteins (Lobry & Gautier, 1994). It is recognized that the degeneracy of the genetic code allows an additional potential for mRNA to carry structural information regarding the encoded protein, which can be at the level of a single codon or at a contiguous nucleotide region (Brunak et al., 1994). The first, second, and third base of the codon have been connected to, respectively, the biosynthetic pathway (Taylor & Coates, 1989), the hydrophobicity pattern (Volkenstein, 1966; Woese et al., 1966), and the helix or beta sheet forming potentiality (Siemion & Siemion, 1994) of the coded amino acid.

To investigate the above issues, we calculated the frequencies (see Materials and methods for details; the calculated values are listed in Table 1) of the 61 sense codons for the 20 amino acids as they occur in mRNAs of highly expressed genes (Sharp & Li, 1987) from *E. coli*. We observed a strong correlation between the sum of such frequencies of the synonymous codons for individual amino acids and the sum of the cellular contents of isoaccepting tRNAs as listed in Ikemura (1985). Regression to a linear function for the amino acids Val, Gly, Ile, Lys, Glu, Asp, Gln, Asn, Tyr, His, and Phe (others were not considered for lack of quantification of the tRNA content) yielded a correlation coefficient of 0.94, with a standard error of 0.008. There is experimental evidence that the total amount of tRNA for a particular amino acid parallels the total usage of that amino acid in proteins for *E. coli* and *Mycoplasma capricolum* (Yamao et al., 1991). Thus, tRNA availability is a strong determinant of the translational speed of the cognate codon, and the cellular levels of tRNAs parallel the codon as well as amino acid usage in highly expressed genes. We therefore used the codon frequency values to quantify the slowness in translating the regions on mRNA.

A set of 37 *E. coli* multidomain proteins with known tertiary structures was identified (Table 2). These proteins generally contained up to four domains, each consisting of several secondary structural elements (helices and β-strands) strongly associated in a globular and presumed independent structural module. The individual domains contained one or two polypeptide segments (*domain-segments*) that are separated in the primary structure of the protein. Domains are demarked either by links that connect peptide segments between two domains or by ends in the case of C-terminal domain-segments. We tested whether these links and ends are encoded by slowly translated regions on mRNA by considering a list containing the len/30 (residue length of protein divided by 30) number of slowest translated nucleotide regions in each of the 37 cases. The len/30 number was chosen because the distribution of lengths for exons corresponding to protein structural or functional folding units peaks at about 32–39 codons; however, a relatively flat peak is maintained up to 60 codons (Dorit et al., 1990). The details of the construction of the len/30 list are given in Materials and methods and are illustrated in Figure 1.

Table 1. *Codon fractional frequency values as calculated from mRNAs of highly expressed genes from Escherichia coli*

Amino acid	Codon	Frequency value[a]	Amino acid	Codon	Frequency value
Met	aug	0.0197	Ala	gcu	0.0470
Trp	ugg	0.0061		gcc	0.0080
Phe	uuu	0.0069		gca	0.0272
	uuc	0.0260		gcg	0.0201
Tyr	uau	0.0055	Val	guu	0.0499
	uac	0.0192		guc	0.0063
Cys	ugu	0.001		gua	0.0227
	ugc	0.0030		gug	0.0120
His	cau	0.0044	Gly	ggu	0.0536
	cac	0.0138		ggc	0.0354
Gln	caa	0.0046		gga	0.0005
	cag	0.0304		ggg	0.0015
Asn	aau	0.0032	Ser	agu	0.0012
				agc	0.0091
Lys	aaa	0.0600		ucu	0.0169
	aag	0.0191		ucc	0.0126
Asp	gau	0.0185		uca	0.0009
	gac	0.0385		ucg	0.0005
Glu	gaa	0.0542	Arg	aga	0.0001
	gag	0.0155		agg	0.0000
Ile	auu	0.0108		cgu	0.0488
	auc	0.0475		cgc	0.0187
	aua	0.0000		cga	0.0003
Pro	ccu	0.0038		cgg	0.0001
	ccc	0.0003		aac	0.0371
	cca	0.0038	Leu	uua	0.0005
	ccg	0.0264		uug	0.0026
Thr	acu	0.0253		cuu	0.0029
	acc	0.0269		cuc	0.0023
	aca	0.0011		cua	0.0003
	acg	0.0034		cug	0.0611

[a] The codon fractional frequency values were calculated as given in Materials and methods.

Table 2. *Proteins, domains, and their constituent domain-segments used in this study*

	PDB/EMBL[a] identifiers	Protein name	Cellular[a] location	Domains and the range[b] of residues forming domain-segments	
TYPE I: Two domains with a single link					
I.1.	3eca-A/ecansba	L-Asparaginase type II	Periplasmic	1-190	213-326
I.2.	1cgp-B/eccrp	Catabolic gene regulatory	Cytosolic	1-129	140-210
I.3.	2trx-A/ecrhoa	Thioredoxin	Cytosolic	1-59	76-108
I.4.	1pii/ectrpc	Bifunctional (isomerase and synthase)	Cytosolic	1-254	256-452
I.5.	1gla-G/ecglyK	Glycerol kinase	Cytosolic	1-252	260-501
I.6.	1fia/ecfis	Factor for inversion simulation	Cytosolic	1-69	74-98
I.7.	1acm-B/ecpyrbia	Aspartate carbamoyl-transferase-regulatory	Cytosolic	1-97	102-152
I.8.	DHDPS/ecdhdps	Dihydrodipicolinate	Cytosolic	1-224	226-292
I.9.	1emd/ecmdh1	Malate dehydrogenase	Cytosolic	1-148	149-312
I.10.	1ctf/ecrpobc	Ribosomal protein L7/12	Cytosolic	1-50[c]	53-120
I.11.	2rsl/ectn1000	Gamma delta resolvase	Cytosolic	1-140[c]	141-183
I.12.	1cdd/ecpurmn	Phosphoribosyl glycin/amide formyltransferase	Cytosolic	1-100	104-212
TYPE II: Two domains with two links					
II.1.	2abk/ecnth	Endonuclease III	Cytosolic	1-16	29-129
				138-211	
II.2.	1dsb/ecdsf	Disulfide oxidoreductase	Periplasmic	1-62	63-138
				139-189	
II.3.	1dra/ecfolx	Dihydrofolate reductase type I	Cytosolic	1-37	38-88
				89-159	
II.4.	1aam/ecaspc	Aspartate aminotransferase	Cytosolic	1-35	36-313
				314-396	
II.5.[d]	1ltp/eclaci	Lactose operon repressor	Cytosolic	62-161	165-305
				308-322[c]	
II.6.	1trb/ectrxb	Thioredoxin reductase	Cytosolic	1-114	115-244
				245-320	
II.7.	1ake/ecadk	Adenylate kinase	Cytosolic	1-121	122-159
				160-214	
II.8.	1goa/ecqrnh	RNase H	Cytosolic	1-69	71-112
				115-155	
II.9.	1acm-A/ecpyrbia	Aspartate carbamoyltransferase-catalytic	Cytosolic	1-136	140-288
				292-310	
TYPE III: Two domains with three links					
III.1.	2liv/eclivhmg	Leu-Ile-Val binding protein	Periplasmic	1-118	124-248
				253-326	331-344[e]
III.2.	2lbp/eclivhmg	Leu binding protein	Periplasmic	1-118	124-248
				253-327	332-346[e]
III.3.	1dmb/ecuw89	D-Maltodextrin binding protein	Periplasmic	1-109	114-257
				262-311	315-370
III.4.	2dri/ecrbsp	D-Ribose binding protein	Periplasmic	1-103	104-235
				236-263	265-271[e]
III.5.	2gbp/ecmglabc	D-Galactose binding protein	Periplasmic	1-110	111-256
				257-293	295-309[e]
III.6.	1abe/ecarafgh	L-Arabinose binding protein	Periplasmic	1-109	109-253
				257-283	287-306
III.7.	1pfk/ecpfka	Phosphofructokinase	Cytosolic	1-124	138-253
				257-300	309-319[e]

(*continued*)

Slow nucleotide regions on mRNA correspond to protein domain links/ends

The results from the search for the slowest mRNA regions (see Materials and methods for definitions) are given in Table 3. A total of 37 proteins with known tertiary structures were examined. They contained 88 domains divided into 112 sequence contiguous domain-segments. The effective number of oligopeptide domain-segment connecting links was 75 and the number of domain ends at the protein C termini was 31, yielding a total of 106 domain links/ends. The analysis shows that the slowest regions on mRNA could often be identified with domain links/ends on proteins and such a correlation is certainly more than expected. Tricodons were considered because three amino acid spans correspond to basic protein folding structural units (see Materials and methods). The percentage of domain links/ends

Table 2. *Continued*

	PDB/EMBL[a] identifiers	Protein name	Cellular[a] location	Domains and the range[b] of residues forming domain-segments			
TYPE IV: Three domains with two links							
IV.1.	1etu/ectgtufb	Elongation factor EF-Tu	Cytosolic	1-198	201-297[c]	301-393[c]	
IV.2.	1kfd/ecpola	DNA polymerase I	Cytosolic	1-323[c]	324-517	521-928	
IV.3.	2pol/ecdnaan	DNA polymerase III beta	Cytosolic	1-109	124-243	254-366	
IV.4.	1bia/ecbira	Biotin operon repressor	Cytosolic	1-60	68-269	274-321	
TYPE V: Three domains with more than two links							
V.1.	3icd/ecicd	Isocitrate dehydrogenase	Cytosolic	1-124 318-416	125-157 203-317	158-202	
V.2.	2glt/ecgshii	Glutathione synthase	Cytosolic	1-121	122-133[e] 208-316	134-201	
V.3.	1pda/echemc	Porphobilinogen deaminase	Cytosolic	1-99 200-217	105-193	222-313	
TYPE VI: Four domains with three links							
VI.1.[d]	1rnr/ecnrda	Ribonucleotide reductase	Cytosolic	1-184	185-291	300-339	340-375[c]
VI.2.[d]	2reb/ecreca	Recombinase A	Cytosolic	1-30	37-268	270-328	329-352[c]

[a] PDB is a database of atomic coordinates of known protein tertiary structures (Bernstein et al., 1977). PDB codes (a number followed by three characters) identify the protein files containing the structural information; an additional letter identifies the particular polypeptide chain used. Data for DHDPS (I.8) is taken from Mirwaldt et al. (1995). For 1dmb (III.3), 1abe (III.6), 1pfk (III.7), 1glt (V.2), and 2pol (IV.3), the demarcation of domains on the proteins was confirmed by viewing the molecular structure. The corresponding nucleotide sequences are as given in the EMBL database (identifiers given after the slash), comprised of the nucleotide sequence of genes of known primary structure (Rice et al., 1993). The information regarding the subcellular location of the protein was taken from the SWISS-PROT protein sequence data bank (Bairoch & Apweiler, 1996). The CAI values (see Materials and methods) of the genes are in increasing order: ECTN1000 (0.18); ECBIRA (0.25); ECLACI (0.31); ECTRPC (0.34); ECHEMC (0.34); ECPYRBIA2 (0.38); ECDHDPS (0.39); ECPURMN (0.40); ECNTH (0.41); ECQRNH (0.42); ECPOLA (0.42); ECPYRBIA1 (0.43); ECLIVHMG1 (0.43); ECGSHII (0.47); ECTRXB (0.47); ECFOLX (0.48); ECLIVHMG2 (0.48); ECDNAAN (0.50); ECGLYK (0.51); ECRBSP (0.51); ECASPC (0.52); ECCRP (0.52); ECARAFGH (0.53); ECNRDA (0.54); ECANSBA (0.57); ECFIS (0.57); ECUW89 (0.60); ECMGLABC (0.60); ECDSF (0.62); ECICD (0.62); ECMDH1 (0.62); ECRHOA (0.65); ECRECA (0.65); ECPFKA (0.72); ECADK (0.73); ECTGTUFB (0.81); ECRPOBC (0.82).

[b] The range of numbered residue portions in the primary structure that form the domain-segments constituting individual domains; they are given by the authors who determined the individual three-dimensional structures (see references in the appropriate PDB files).

[c] Domain boundaries are known (see references in PDB), but the tertiary structure for these domains (seven in number) has not yet been determined.

[d] For 1ltp (II.5), data corresponds to the repressor core according to the model of Matthews and coworkers (Nichols et al., 1993). However, the structure of the complete protein as determined by X-ray crystallography appeared shortly after manuscript submission (Lewis et al., 1996; Matthews, 1996). Discussion pertaining to the new structure is included in the text (see the Discussion, "Mutation data in support of the hypothesis"). For 1rnr (VI.1) and 2reb (VI.2), the regions corresponding to 340-375 and 329-352, respectively, were not visible in the electron density map and were presumed to be disordered. We have assigned them tentatively as the respective fourth domains.

[e] The domain-segments are of short length (≤15 amino acids), of which six are C-terminal (ends) and one is domain-segment internal.

represented by slowly translated codons at or within 14 residues from them (the "*space*" parameter as defined in Materials and methods) with statistical significance at two standard deviations (σ) is 67 (Table 3); 97% of the proteins considered possessed at least one link/end identified with a slow region. The expected frequency was based on the total length of the proteins considered and that of the links/ends, along with the ±14 residue spans flanking them. The average length of a domain is 132 residues and that of a domain-segment is 98 residues, considerably greater than 14. A statistical significance of 3σ is attained when the search is restricted to a span of ten residues flanking the links/ends, representing still 60% of the links/ends and 97% of the proteins (Table 3). The statistical significance increases to 6.2σ as the spacing from the links/ends considered is reduced to zero, such that the slowest coding regions are considered only when they fall within the structurally defined protein domain links and C-terminal ends; and yet, the percentage of proteins for which at least one link/end is represented is not considerably reduced. For example, 86% of the proteins are still represented at a "*space*" level of four residues.

The level of significance increases at every spacing (Table 3) because only the very slowest tricodons are considered; i.e., as *n* increases in *len/n*, where *len* corresponds to the protein residue length and *n* to the expected domain-segment (exon) length, and the number of slowest tricodons taken for each protein is *len/n*. This observation shows that the regions on mRNA representing the links/ends on proteins are among the slowest possible. Thus, a pause induced by slow translation of the regions is likely to occur in the growth of the peptide at domain boundaries.

Positional distribution of the slow regions with reference to domain links

The positional distribution of the slow regions as found in domain links (Table 4) indicates that for higher "*space*" groupings, the slow regions are preferentially 3′ to link spans on mRNA.

A **Protein entry : 1goa/ecqrnh (II.8 in Table 2).**

Length = 155 residues; 2 domains with 2 inter-domain links and one end (for the C-terminal domain).

One domain is delineated by residues 1-69 and 115-155 while the other domain contains 71-112. The range of residues forming the links and ends are determined as given in Materials and methods section. The links are a = 70-72 and b = 113-115 while the end is c = 153-155.

Tricodon locations (and the sum of frequency values of the constituent codons) arranged in decreasing order of slowness are as follows:

```
143 (.005);  18 (.006);  16 (.008);  124 (.013);  103 (.013);  66 (.015);
 65 (.020); 125 (.022); 142 (.023);   28 (.025);   26 (.025); 102 (.027);
101 (.027); 100 (.028);  25 (.030);   49 (.030);  ----------   86 (.169).
```

The len/30 list will contain 5 (= 155/30) slowest regions from the above and the list is as follows:

	Slow region		Link/end	Space
1.	143,142	c	(153-155)	-10
2.	124,125	b	(113-115)	+11
3.	103	b	(113-115)	-10
4.	66,65	a	(70-72)	-4
5.	28		---	--

The list gives the following values :
nre = 4; ndmr = 4; l/e = 3

B **Protein entry : 1cdd/ecpurmn (I.12 in Table 2).**

Length = 212 residues; 2 domains with one inter-domain link and one end (for the C-terminal domain).

N-terminal domain is delineated by residues 1-100 while the C-terminal domain is delineated by 104-212. The link a = 101-103 and the end b = 210-212.

Tricodon locations (and the sum of frequency values of the constituent codons) arranged in decreasing order of slowness are as follows:

```
116 (.001);  12 (.005);  11 (.005); 115 (.006);  54 (.010);  53 (.011);
134 (.014); 117 (.015); 100 (.015); 135 (.017);  71 (.018);  69 (.018);
176 (.020); 183 (.021); 182 (.022);  91 (.022); ---------- 188 (.170).
```

The len/30 list will contain 7 (= 212/30) slowest regions from the above and the list is as follows:

	Slow region		Link/end	Space
1.	116,115,117	a	(101-103)	+15
2.	54,53		---	--
3.	134,135		---	--
4.	100	a	(101-103)	-1
5.	71		---	--
6.	176		---	--
7.	183		---	--

The list gives the following values:
nre = 4; ndmr = 2; l/e = 1

Fig. 1. A: Illustration of the creation of an exemplary len/30 list of slowest mRNA regions. *Space* indicates the nearness of a slow region to the codons expressing residues within protein domain links/ends, as defined in Materials and methods. The *space* is not listed when the slow region is located further than 16 amino acids from the terminii of the link/end. The term *ndmr* represents the number of slow regions that are near to the domain links/ends by a maximum space of 16 codons; *nre* represents the number of slowest regions required to be considered from the len/30 list to observe the *ndmr* regions; and *l/e* represents the number of links and ends represented. **B:** Derivation of the *nre* value for a protein where *nre* is greater than *ndmr*. Tricodon locations at 18 and 16 for A and 12 and 11 for B were ignored because they fall within the first 20 codons (see Materials and methods).

Table 3. *Results of the search for slow regions on mRNA that code for domain links and ends*

Space[b] (±)	(len/30) list[a] *nre*[c]	*ndmr*[d]	*o* − *e*[e] (σ)	*l/e*[f] No.	*l/e*[f] %	*np*[g] No.	*np*[g] %	(len/40) list *o* − *e* (σ)	*l/e* %	*np* %	(len/50) list *o* − *e* (σ)	*l/e* %	*np* %	(len/60) list *o* − *e* (σ)	*l/e* %	*np* %
16	296	103	1.2	74	70	37	100	2.5	59	95	3.0	47	78	3.2	44	76
14	284	95	2.0	71	67	36	97	3.1	57	92	3.2	45	78	3.3	42	76
12	275	84	2.2	66	62	36	97	3.5	53	92	3.9	42	76	4.1	40	70
10	263	76	3.0	64	60	36	97	3.7	50	92	4.0	41	76	4.3	37	70
8	246	62	3.0	54	51	35	95	3.5	42	89	3.4	31	70	3.6	27	62
6	232	53	3.7	48	45	35	95	4.6	39	86	4.8	27	65	5.4	24	57
4	218	41	3.7	41	39	32	86	4.5	30	76	5.0	21	54	5.4	17	46
2	158	26	4.3	26	25	21	57	5.2	18	46	5.2	14	38	5.6	10	27
0	70	13	6.2	13	12	12	32	7.4	8	22	7.1	8	19	8.1	6	14

[a] (len/*n*), number (residue length of protein divided by *n*) of the slowest mRNA regions used in this study for each of the 37 proteins (see Materials and methods for details on this and the following footnotes).

[b] Space (defined in the legend to Fig. 1), nearness of a slow region to the actual link/end in terms of number of codons on either side separating the slow region from the link/end terminii.

[c] *nre*, a subset of the len/*n* list of regions for all the 37 proteins (illustrated in Fig. 1); total number of distinct slow regions required to yield *ndmr* regions that are near to the links/ends within a given space.

[d] *ndmr*, number of contiguous slowest codon spans that occur on mRNA near or at regions corresponding to the protein domain links/ends, whereas *nre* is the number of slowest regions required to be considered from the len/*n* list to observe the *ndmr* regions. The *nre* and *ndmr* values for each protein are summed.

[e] *o* − *e*, difference in the observed and expected number of *ndmr* regions in terms of standard deviations (σ).

[f] *l/e*, absolute number (No.) and percentage (%) of the total number of links/ends (the 37 proteins that are considered in this study contain 106 links/ends) that are represented by *ndmr* regions.

[g] *np*, absolute number (No.) and percentage (%) of the total number of proteins (37) considered in this study for which at least one link/end is represented by a *ndmr* region.

No such bias is shown in lower "*space*" groupings. This preferential tendency of the ribosome to pause after (or nearly so) it has translated a domain/domain-segment supports further the relationship between pause and domain folding. It may be noted here that based on observations in two (*aro1* and partly in *fas1*) of five genes considered from yeast, Brown and coworkers (Purvis et al., 1987; McNally et al., 1989) could suggest the possibility of the existence of such domain correlation for proteins that do not refold in vitro to their native conformation. Krasheninnikov et al. (1989) demonstrated that rare codon clusters can determine the boundaries of polypeptide chain fragments of the same secondary structural type.

Slow regions on the nascent peptide and the peptide channel on the ribosome

The preference/avoidance of the 20 amino acid types in slow (*D*) regions as compared with their general occurrence (*T*) in the 37 proteins considered in the study or in the search (*S*) protein regions that encompass the links/ends along with the adjacent re-

Table 4. *Positional distribution of the slowest mRNA regions as they pertain to only the domain links*

n in (len/*n*)[a] list	No. of regions falling within the N- or C-terminal sides of the links: Lower space groupings[b] 1-6 N_T,C_T	1-8 N_T,C_T	1-9 N_T,C_T	1-10 N_T,C_T	Higher space groupings[b] 7-16 N_T,C_T	9-16 N_T,C_T	10-16 N_T,C_T	11-16 N_T,C_T
30	17, 17	19, 20	20, 22	23, 25	14, 24	12, 21	11, 19	8, 16
40	13, 15	15, 17	16, 19	17, 22	10, 19	8, 17	7, 15	6, 12
50	9, 8	10, 10	11, 12	12, 15	4, 17	3, 15	2, 13	1, 10
60	8, 7	9, 8	10, 10	11, 13	4, 15	3, 14	2, 12	1, 9

[a] The (*len/n*) list contains a given number (residue length, *len*, of protein divided by *n*) of slowest regions in each of the 37 proteins considered in this study.

[b] Slow regions are counted if they occur within or at either side of the tricodons expressing residues for the protein domain links within the maximum codon spacing listed; N_T, toward the N-terminal side; C_T, toward the C-terminal side of the polypeptide link.

gions with a given "*space*" was examined. The results are given in Table 5 under the columns designated as *D/S* and *D/T*. The analysis of the values for *D/S* and *D/T* indicated that the amino acids preferred in slow regions that occur at or near the domain links/ends are Tyr, His, Trp (large hydrophobic aromatic residues); Ile, Leu, Val (hydrophobic and branched); Ser, Thr [side chains with a hydroxyl group that prefers to H-bond to main-chain atoms (Bordo & Argos, 1994)]; Pro (hydrophobic, cyclic, and exerts significant constraint on the conformation of the backbone); and Cys (hydrophobic and sulfated side chain). Charged and polar amino acids are generally not preferred. Glycine, which provides opportunity for large conformational flexibility in the main chain, is also avoided. The propensity for proline and the avoidance of glycine imply a preference for constraints in the backbone conformation.

The preferred amino acids coded in slow regions are almost all decidedly hydrophobic, perhaps used to "stick" to the ribosome cavity also considered hydrophobic (van-den-Broek et al., 1989; Kim et al., 1991; Hardesty et al., 1992). It was reported earlier (Picking et al., 1991) that a synthesized hydrophobic polypeptide appeared to accumulate as a hydrophobic mass adjacent to the peptidyl transferase center on the ribosome, whereas polylysine extended directly from the ribosome into the surrounding solution. The regions encompassing hydrophobic residues may thus by interaction with the cavity cause an additional pause during the exit of the nascent peptide to the cytoplasm. Interaction between nascent protein chain and ribosome has been suggested as an explanation for the the increase in the stability of the translation complex as a function of the distance translated on mRNA (Goldman et al., 1995). Further, the translation of hydrophobic segments on the nascent peptide causes a pause by the ribosome on the mRNA (Kim et al., 1991).

The average solvent-accessible surface exposure (see Materials and methods) of the specific 20 residue types forming slow regions (indicated by E_D in Table 5) was compared with the expected level based on their mean exposure throughout the pro-

Table 5. *Characterization of the residues in protein regions corresponding to slow translation spans on mRNA*[a]

	Percentage of usage in			Ratios of usage			Exposure to solvent in		Ratio of exposure
	Slow regions D[b]	Search regions S[c]	Proteins considered T[d]	D/S	D/T	D/R[e]	Proteins considered E_T[f]	Slow regions E_D[g]	E_D/E_T
Cys	3.96	1.16	0.95	3.41	4.17	8.69	9.7	5.9	0.61
His	3.96	1.77	1.78	2.24	2.22	2.20	55.9	29.2	0.52
Ser	10.32	4.55	4.72	2.27	2.19	2.50	36.1	28.5	0.79
Pro	8.33	4.39	4.08	1.90	2.04	2.42	48.1	53.5	1.11
Tyr	5.16	3.12	2.87	1.65	1.80	2.11	43.1	39.5	0.92
Trp	1.58	1.08	1.03	1.46	1.53	2.62	37.6	21.5	0.57
Ile	7.14	5.32	6.13	1.34	1.16	1.22	19.7	21.6	1.10
Val	8.33	7.20	7.59	1.16	1.10	0.91	19.0	29.4	1.55
Leu	11.90	10.44	9.29	1.14	1.28	1.71	20.3	24.7	1.22
Thr	5.95	5.39	5.32	1.10	1.12	1.06	32.4	31.5	0.97
Met	1.98	2.00	2.34	0.99	0.85	1.02	30.3	45.8	1.51
Phe	2.78	2.93	3.07	0.95	0.91	0.85	26.4	21.8	0.83
Arg	3.97	4.97	4.82	0.80	0.82	0.59	81.3	97.3	1.20
Ala	7.14	8.98	10.21	0.80	0.70	0.69	23.6	28.4	1.20
Gln	3.17	4.43	4.22	0.72	0.75	0.90	76.3	47.0	0.62
Asp	4.37	6.55	6.26	0.67	0.70	0.77	65.9	65.9	1.00
Asn	2.78	4.35	4.13	0.64	0.67	0.69	62.9	71.6	1.14
Gly	3.97	7.59	8.12	0.52	0.49	0.44	24.0	24.1	1.00
Glu	2.38	6.86	7.02	0.35	0.34	0.34	86.1	99.6	1.16
Lys	0.79	6.93	6.04	0.11	0.13	0.10	100.9	98.0	0.97

[a] Data given in the table was calculated using the *ndmr* regions (Table 3) for len/30 list at a *space* of ±12. A *space* of 12 was chosen because more than 60% of links/ends are represented at this space with a statistical significance > 2.0. The conclusions given by the presented data in the table did not change with *spacing*.

[b] *D*, Percentage of usage in *ndmr* regions (these regions are slow and occurred within the search regions where search region includes codons expressing links/ends as well as the adjacent regions within ±12 residues in the case of links and −12 in the case of ends).

[c] *S*, percentage of usage in search regions.

[d] *T*, percentage of usage in all the proteins considered in this study (Table 2).

[e] *D/R*, ratio between the percentages of amino acid usage in slow regions and in highly expressed proteins. A set of 60 highly expressed proteins corresponding to the set of highly expressed genes (see Materials and methods) was used to calculate the percentage of amino acid usage.

[f] E_T, average solvent accessibility surface (Å^2) (see Materials and methods) of individual residues as they occur in all the proteins considered in this study.

[g] E_D, average solvent accessibility surface (Å^2) (see Materials and methods) of individual residues as they occur in *ndmr* regions mentioned above.

tein structure (indicated by E_T in Table 5). Of the five preferred nonpolar hydrophobic amino acids, four (Pro, Ile, Val, Leu) showed enhanced exposure given that the ratio of E_D to E_T was >1.0. Similarly, the four preferred polar but uncharged residues (Cys, Ser, Thr, and Tyr) showed a reduced surface accessibility, as was also the case with the preferred charged residue His. These results could suggest that the residues forming the slow regions are not essential for the final domain structure in proteins because they are unlikely, for example, to act as particularly buried hydrophobic nucleation centers or as greatly exposed polar residues maintaining solvent association to aid the folding process (Bowie & Sauer, 1989; Bowie et al., 1990; Yue & Dill, 1992; Kamtekar et al., 1993). As a result, they can remain in the ribosome channel and thus indirectly foster a translational pause that directly aids domain folding.

The amino acid usage in slow regions was compared with that in the extended linker regions (D/S) as well as with that in the proteins of the data set (D/T) under study because it is possible that the slow region biases may be due to the distinct functionalities of different domains and/or linkers. It is widely accepted that most multidomain proteins have resulted from fusion of genes coding for individual domains (Edelman et al., 1969; Rossmann et al., 1974). The domain linkers are oligopeptides that connect domains through gene fusions and such linkers were found to be devoid of hydrophobic residues; the constituent amino acids in the linkers also showed enhanced exposure in the native structure (Argos, 1990). However, it has been shown here that the slow regions are rich in hydrophobic residues and no preferential burying or exposure occurred for these regions in the native protein folds. Further, we subjected the correlations between D/S and S, D/T and T, and D/R and R (Table 5) to linear regression analysis and obtained correlation coefficients -0.53, -0.55, and -0.58 and corresponding slopes -0.16, -0.19, and -0.38 for the three cases, respectively. The higher slope found for the correlation with the composition in the highly expressed genes (D/R) compared with that found with linker regions or the proteins in the data set indicate that the amino acid composition bias in the slow regions is driven mostly, but still partially, by the rarity or abundance of residue types in the highly expressed proteins rather than by the linkers or the domains. It must be emphasized that the correlations are nonetheless small and thus the codon frequency values are effectively not responsible for the composition patterns observed here in slow regions. Thus, the observed bias in the amino acid composition in slow regions is not due to the characteristics of linkers or domains, or even evolutionary vestiges of translation signals from the ancestral single-domain proteins that have fused to yield multidomain proteins.

Synonymous codon usage in the slow mRNA regions

Because codon frequency values (which reflect both synonymous codon usage as well as amino acid usage) were used as a measure of translation speed, which reflects biases in synonymous codon usage and in amino acid composition, synonymous codon usage in slow regions was compared with that generally found in the mRNAs of highly expressed genes. The results are given in Table 6. The values in the columns designated *srx* and *sdx* indicate that for 51 of the 61 sense codons there is a change in the sign of *srx* (+ sign indicates preference and − sign indicates avoidance of codon in highly expressed genes) and *sdx* (similarly for slow regions) values or *sdx* approaches zero (see Materials and methods for definitions). This implies that the synonymous codons preferred in highly expressed genes are generally avoided in slow regions and those avoided in highly expressed genes are generally preferred in slow regions, in agreement with the rationale used in this study to quantify slow regions in mRNAs. The observation points out that, although the codon frequency value has been used, the synonymous codon usage is in agreement with the accepted protocols in codon usage studies. In ten exceptional cases, the *srx* and *sdx* values are of the same sign (indicated by # in Table 6) and thus share similar preference or avoidance in highly expressed genes and in slowly translated mRNA sequences. Two (ugc and cgc) of the ten showed preference in both highly expressed genes and also in slow regions, whereas the remaining eight showed avoidance in both highly expressed genes and in slow regions. Six of these eight cases are rare codons. Rare codons (as defined in Materials and methods and indicated by * in Table 6) are 12 in number. Bonekamp et al. (1989) and Curran and Yarus (1989) observed that, whereas different codons are translated at different rates and abundant codons are mostly translated at a higher rate, certain rare codons are also translated at an unexpectedly higher rate. It is noteworthy that at least four of the eight avoided codons in both highly expressed genes and in slow regions belong to this latter category of rare codons.

An assessment of how different the synonymous codon usage in slow regions is from that in other regions of the genes was done by calculating the synonymous codon frequency in the "search" regions (linkers along with 12 codons on either side) and *slx* (similar to the above-mentioned *srx* and *sdx* values, but now dependent on the codon usage in the search regions). Comparison of *sdx* (domain slow regions) with *slx* (search regions) largely indicated a pattern very similar to that of *sdx* compared with *srx* (highly expressed genes) except in the codon auu for Ile. The synonymous codon frequency in the data set of 37 proteins were not significantly different from that in the search regions (data not shown). Thus, codon usage in slow regions is significantly different from codon usage in other regions of the mRNA.

Levels of expressivity of the 37 genes considered in the study

The codon adaptation index (CAI) as formulated by Sharp and Li (1987) was used (see Materials and methods) as an indication of the expression level of each of the 37 genes considered in this study. The resulting CAI values (Table 2 legend) indicated that the genes studied are of mixed expression levels with the overall average CAI at 0.51, the average for the highly expressed genes at 0.97, and the mean for lowly expressed genes (*trpr*, *dnag*, *eltA*, and *galR*) at 0.21. The detected slow regions in a gene are relative to the overall codon frequency values of other regions in the same mRNA.

Codon usage and translation rates

Codon usage in slow regions can also be compared with observed translation rates (given in the final column of Table 6) where known (Curran & Yarus, 1989). It is noteworthy that in six (Phe, Tyr, Gln, Pro, Ser, Arg) of nine amino acid types for which the translation speed is known for the synonymous co-

Table 6. *Characterization of synonymous codon (syncod) usage in the slow regions on mRNA compared with that in highly expressed genes (HEG)*[a]

			Syncod frequency in					
	Codon[b]	*RSCU*[c]	HEG mRNA *sr*[d]	Slow regions *sd*[e]	*slx*[f]	*srx*[f]	*sdx*[f]	Speed[g]
Met	aug							
Trp	ugg							
Phe	uuu	0.42	0.2109	0.7143	−0.11	−0.29	0.21	8.5
	uuc	1.58	0.7891	0.2857	0.10	0.29	−0.22	12.0
Tyr	uau	0.44	0.2213	0.6154	−0.04	−0.28	0.12	4.3
	uac	1.56	0.7787	0.3846	0.03	0.28	−0.12	8.4
Cys	ugu	0.68	0.3409	0.3000	−0.07	−0.16	−0.20#	4.0
	ugc	1.32	0.6591	0.7000	0.07	0.16	0.20#	7.0
His	cau	0.49	0.2428	0.5000	−0.07	−0.26	0.00	4.0
	cac	1.51	0.7572	0.5000	0.07	0.26	0.00	8.0
Gln	caa	0.26	0.1321	1.0000	−0.25	−0.37	0.50	5.6
	cag	1.74	0.8679	0.0000	0.25	0.37	−0.50	10.0
Asn	aau	0.16	0.0783	1.0000	−0.22	−0.42	0.50	
	aac	1.84	0.9217	0.0000	0.22	0.42	−0.50	
Lys	aaa	1.52	0.7583	0.0000	0.31	0.26	−0.50	
	aag	0.48	0.2417	1.0000	−0.31	−0.26	0.50	
Asp	gau	0.65	0.3247	0.8182	0.03	−0.18	0.32	
	gac	1.35	0.6753	0.1818	−0.03	0.18	−0.32	
Glu	gaa	1.56	0.7782	0.0000	0.22	0.28	−0.50	
	gag	0.44	0.2217	1.0000	−0.22	−0.28	0.50	
Ile	auu	0.56	0.1856	0.9400	0.12	−0.15	0.61	
	auc	2.44	0.8144	0.0000	0.17	0.48	−0.33	
	aua*	0.00	0.0000	0.0600	−0.30	−0.33	−0.27#	
Pro	ccu	0.44	0.1104	0.1428	−0.15	−0.14	−0.11#	8.4
	ccc*	0.04	0.0092	0.1428	−0.14	−0.24	−0.11#	9.6
	cca	0.44	0.1104	0.4286	−0.05	−0.14	0.18	1.6
	ccg	3.08	0.7699	0.2857	0.35	0.52	0.03	2.5
Thr	acu	1.79	0.4471	0.1333	0.01	0.20	−0.12	
	acc	1.90	0.4750	0.0000	0.22	0.23	−0.25	
	aca*	0.07	0.0186	0.1333	−0.20	−0.23	−0.12#	
	acg	0.24	0.0594	0.7333	−0.04	−0.19	0.48	
Ala	gcu	1.84	0.4594	0.0000	−0.03	0.21	−0.25	
	gcc	0.31	0.0781	0.7778	−0.01	−0.17	0.53	
	gca	1.06	0.2662	0.0000	−0.02	0.17	−0.25	
	gcg	0.79	0.1963	0.2222	0.05	−0.05	−0.03	
Val	guu	2.17	0.5491	0.0000	0.12	0.30	−0.25	
	guc	0.28	0.0694	0.1905	−0.09	−0.18	−0.06	
	gua	1.00	0.2497	0.1429	−0.07	0.00	0.11	
	gug	0.53	0.1318	0.6667	0.03	−0.12	0.42	
Gly	ggu	2.36	0.5889	0.0000	0.15	0.34	−0.25	
	ggc	1.56	0.3891	0.0000	0.21	0.14	−0.25	
	gga*	0.02	0.0058	0.4000	−0.20	−0.25	0.15	
	ggg*	0.06	0.0162	0.6000	−0.15	−0.23	0.35	
Ser	agu	0.17	0.028	0.1923	−0.07	−0.14	0.03	
	agc	1.33	0.2214	0.1538	0.08	0.05	−0.01	
	ucu	2.46	0.409	0.0385	0.07	0.24	−0.13	11.6
	ucc	1.83	0.305	0.1154	0.06	0.14	−0.05	14.7
	uca	0.14	0.0229	0.1923	−0.07	−0.14	0.03	7.0
	ucg*	0.08	0.0127	0.3077	−0.06	−0.15	0.14	9.0

(*continued*)

dons, the codons that are translated with faster speeds compared with their synonymous codons are avoided in slow regions; i.e., *sdx* is negative. The exceptions make codons for Cys, His, and Leu. However, cug (Leu), which displays the highest translation speed (14.4) compared with its synonymous codons, is avoided in slow regions. In the case of His, the codon usage in slow regions has become neutral. Thus, there is strong agreement between codon usage in slow regions and reported codon translation rates such that slow regions as observed in this work might well bring about the required time pauses.

Table 6. *Continued*

	Codon[b]	*RSCU*[c]	Syncod frequency in HEG mRNA *sr*[d]	Slow regions *sd*[e]	*slx*[f]	*srx*[f]	*sdx*[f]	Speed[g]
Arg	aga*	0.01	0.0015	0.3000	−0.13	−0.16	0.13	
	agg*	0.00	0.0000	0.0000	−0.17	−0.17	−0.17#	
	cgu	4.30	0.7172	0.0000	0.31	0.55	−0.17	14.0
	cgc	1.65	0.2751	0.4000	0.28	0.11	0.23#	9.0
	cga*	0.03	0.0046	0.0000	−0.15	−0.16	−0.17#	11.5
	cgg*	0.01	0.0015	0.3000	−0.14	−0.17	0.13	0.8
Leu	uua*	0.05	0.0075	0.1333	−0.10	−0.16	−0.03	4.3
	uug	0.23	0.0377	0.3333	−0.06	−0.13	0.17	8.7
	cuu	0.25	0.0422	0.2333	−0.10	−0.12	0.07	8.4
	cuc	0.20	0.0331	0.2667	−0.09	−0.13	0.10	11.0
	cua*	0.03	0.0045	0.0333	−0.15	−0.16	−0.13#	0.6
	cug	5.25	0.8750	0.0000	0.50	0.71	−0.17	14.4

[a] Results were calculated using *ndmr* regions for the len/30 list with space ± 12.
[b] Twelve *rare* codons (see Materials and methods for details) are marked with (*).
[c] *RSCU* (relative synonymous codon usage) values (as defined in Materials and methods) were calculated here with the data set of mRNAs from highly expressed genes.
[d] *sr*, synonymous codon frequency as calculated from the mRNAs of highly expressed genes.
[e] *sd*, synonymous codon frequency as calculated using the *ndmr* slow regions.
[f] *slx*, *srx*, and *sdx* are as defined in Materials and methods. The parameters allow distinction in the preference of one codon over the other synonymous codons and comparison between codon selection in slow regions (*srx*) to that (*sdx*) in highly expressed genes and also (*slx*) in search regions. A negative value indicates avoidance of the codon in the given regions or genes, whereas a positive value denotes preference. The codon cases where the values of *srx* and *sdx* have not involved a change in sign are indicated by (#).
[g] Translation speed of the codon as reported in Curran and Yarus (1989). Values are given only for those codons for which the translation rates have been measured in the cited paper. These values, denoted as R_{tRNA}/R_{shift} in Curran and Yarus (1989), were measured by placing aminoacyl-tRNA selection at individual codons in competition with a frameshift assumed to have a uniform rate and they indicate the relative rates of association between the codon and the cognate aminoacyl-tRNA.

Rare codons and their usage in slow regions

Curran and Yarus (1989) and Bonekamp et al. (1989) report that certain rare codons are translated rapidly. In this work, 12 rare codons have been defined (Table 6) and they often overlap with those studied experimentally. It would be of interest to analyze the usage of some of these rare codons in slow regions given their reported high translation rates.

The 12 rare codons can be classified into three groups. The first group consists of **rare codons that are translated rapidly** and includes aua [speed = fast (Bonekamp et al., 1989); *sdx* = −2.7]; cga (speed = 11.5, *sdx* = −0.17); uua (speed = 4.3, *sdx* = −0.03); ccc (speed = 9.6, *sdx* = −0.11); and ucg (speed = 9.0; *sdx* = 0.14). A positive (negative) *sdx* value indicates usage in slow regions more (less) than expected (see Materials and methods for definitions). Thus, in four of the five cases of rare codons that possess an unexpectedly high translation rate, the codon is avoided in slow regions. This observation demonstrates that, though our rationale in searching for slow regions is based on the occurrence of codons in highly expressed genes, the slow regions are also consistent with the experimental data on the rate of translation of rare codons.

The second group is represented by **rare codons that are translated very slowly** and includes cua (speed = 0.6, *sdx* = −0.13); agg [speed = very slow relative to that of cua for leucine (Curran & Yarus, 1989; Bonekamp et al., 1989), *sdx* = −0.17], and cgg (speed = 0.8, *sdx* = .13). Two of these three rare codons that are translated very slowly are avoided in slow regions, contrary to expectation that rare and slowly translated codons should be overrepresented in the slow regions. However, we propose that these two slowly translated codons are avoided by slow regions (that occur at mRNA spans that correspond to the strategic positions on proteins such as domain ends) because they may induce dissociation of the translation complex and cause a deleterious rather an intentional pause. Many reports in the literature (Robinson et al., 1984; Bonekamp et al., 1985; Spanjaard & van Duin, 1988; Varenne et al., 1989, Chen & Inouye, 1990; Spanjaard et al., 1990; Kinnaird et al., 1991; Rosenberg et al., 1993; Goldman et al., 1995) show that the introduction of cua (Leu) or agg (Arg) codon strings results in the reduction of protein yield and presumably induces dissociation of the translation complex or frameshifting and hopping (Kane et al., 1992) events. Chen and Inouye (1990) observed preferential use of such codons in the region spanning the first 25 codons of the mRNA; they suggest that such a preference for the minor codons in an early gene section may modulate gene expression by premature termination of translation, thereby avoiding unnecessary translation of a large part of the mRNA.

The final group of rare codons includes **those for which the data on their translation rate is not available**; namely, aca (*sdx* = −0.12); gga (*sdx* = 0.15); ggg (*sdx* = 0.35); and aga (*sdx* = 0.13). A check for consistency of the preference/avoidance of these codons in slow mRNA regions defined in this study and translation rate must await further experimental results.

Involvement of slow nucleotide regions in mRNA secondary structure

It has been shown that the ribosome pauses directly over a frameshift sequence and such a pause is aided by a higher-order structural element present on the mRNA five nucleotides downstream of the frameshift sequence (Tu et al., 1992). Occurrence of small hairpin stem (in which RNA loops back to form base paired interactions with a 5′ region) structures in the immediate positions downstream of codons has been hypothesized to increase the time of translocation from the A to P site on the ribosome (Shpaer, 1985). The stem-loop structures on mRNA have been suggested as stimulators of ribosome pause to effect not only programmed translational frameshifts, but also other unusual "recoding" events during translation such as selenocysteine incorporation, ribosome hopping on mRNA, and readthrough of translation-termination signals (Gesteland et al., 1992; Farabaugh, 1996). In this study, we find that domain ends are preferentially coded by translationally slow mRNA regions and it would be noteworthy if such slow regions were also involved in secondary structure formation. For this purpose, we considered only the nucleotide regions involving the slow codons that delineate interdomain links on the proteins and are characterized by ± 12 *space*. When more than one slow region is associated with a link, the slow region with a positive *space* value was considered with a further preference for the nearest slow region. Thus, of the 75 interdomain links considered, 49 were identified with a slow region and for each the ability of the corresponding mRNA fragment to form small hairpin stem structures was scrutinized by considering a nucleotide fragment of length 30 nucleotides inclusive of those of the slow codons. The results of the search (see Materials and methods) for secondary structure are presented in Table 7 and summarized here.

Hairpin stem structures could be formed in all cases but one [ECGLYK (I.5) in Table 2]. In 36 of the 48 cases, the stems involved nucleotides from the slow codons; in 9 of the remaining 12, the stem is located at a maximum of 5 nucleotides downstream from the slow codon, whereas in 3, it was a maximum of 9 nucleotides. The distribution of loop length and number of such hairpins given in parentheses was 2 (3), 3 (12), 4 (9), 5 (13), 6 (6), 7 (2), 8 (1), and 11 (2). Thus, in 39 cases, the loop length was in the typical range of 3–6 nucleotides; in one of the two extreme cases of loop length of 11 nucleotides, the hairpin stem is of 7 base pairs; in two of the three cases of loops with 2 nucleotides, all 4 base pairs are of Watson–Crick (WC) type. In 43 of the 48 examples, the stem is of length 4 or more base pairs with distribution of stem length (number of stems) 4 (31), 5 (9), 6 (2), and 7 (1). The remaining five hairpin stems are of 3 base pairs; in all these 5 cases, the stems can be made of 4 base pairs, leaving a short hairpin loop. The total length of stems is 203 (at an average of 4.23 base pairs per stem) and the total length of the loops is 223 (at an average of 4.64 nucleotides). Of the 203 base pairs, 165 (81%) are of WC type. The 38 noncanonical base pairs are distributed as A.G (8), G.A (10), U.G (8), and G.U (12). The 165 WC base pairs are distributed as 88 G-C and 77 A-U. Thus, the data illustrate that the slow regions can form secondary structures that might provide an additional time pause (Shpaer, 1985; Gesteland et al., 1992; Tu et al., 1992; Farabaugh, 1996) for the movement of the ribosome on the mRNA at the critical regions that encode protein domain boundaries.

Implications for studies on codon usage

The correlation between codon usage and tRNA content has been observed extensively (Chavancy et al., 1979; Post & Nomura, 1980; Bennetzen & Hall, 1982; Gouy & Gautier, 1982; Ikemura & Ozeki, 1983; Ikemura, 1985; Sharp et al., 1986). Codon usage has been used to detect coding regions in DNA (Ficket, 1982; Staden & McLachlan, 1982), to predict the levels of protein production from genes (Gribskov et al., 1984; Sharp & Li, 1987), and to suggest which codons are optimal for the translation system of a given organism (Ikemura & Ozeki, 1983; Ikemura, 1981). The occurrence of slow regions in mRNA spans corresponding to domain links/ends is shown in this report to be statistically significant. This correlation represents yet a further advancement in the studies on codon usage and the functional manipulation of codon discrimination in the transfer of genetic code information to protein structure.

Cytosolic versus periplasmic proteins

Proteins can be categorized as cytosolic or noncytosolic. In *E. coli*, the organism from which the proteins considered for this study are derived, the three noncytosolic locations are inner membrane, periplasmic space, and outer membrane. The only protein that is secreted across both the membranes is hemolysin by enterobacteria. Thus, the only targeting in *E. coli* is export to the periplasm and insertion into the cell membrane; eukaryotic cells with their cellular compartments or organelles involve much more export. The correlation between mRNA slow regions and protein domain boundaries in periplasmic versus cytoplasmic proteins were examined as folding principles could be different for the two cases.

As noted in Table 2, 8 of the 37 proteins considered in this study are periplasmic. They are L-asparaginase type II (I.1 in Table 2), disulfide oxidoreductase (II.2 in Table 2), and the amino acid, sugar transport proteins (III.1–III.6 in Table 2). The 8 periplasmic proteins comprise 21 links and 4 C-terminal ends, whereas the cytosolic proteins comprise 54 links and 27 C-terminal ends. The effective total residue length of these 8 periplasmic proteins is 2,301, whereas that of the 29 cytosolic proteins is 8,184. The effective total residue length of the links/ends for the periplasmic proteins is 104, whereas that of the cytosolic proteins is 320.

The statistical significance (calculated as described in Materials and methods, but with the above values for the parameters) of the domain links/ends being represented by slow regions in cytosolic proteins is compared with that in periplasmic proteins (Table 8). The significance is higher in the case of cytosolic proteins for each *len/n* list and *space* value tested. However, the percentage of links/ends represented with a statistical significance of 2σ or more is not much different for both cases of proteins; for example, in the len/40 list, 57% of cytosolic links/ends are represented with a statistical significance of 2.4σ, whereas the percentage for periplasmic case is 56% with significance 2.0σ. Similar observations were obtained for the set of proteins (data not shown) for which at least one link/end is represented by a slow region. Although the correlation between slow mRNA regions and protein domain boundaries is more emphatic in cytosolic proteins, such a correlation cannot be dismissed for periplasmic proteins. The agreement between the results for the two protein classes is also supported by the fol-

Table 7. *Potentiality of the translationally slow nucleotide regions to form hairpin secondary structures*[a]

Table 2 ID	PDB/EMBL	Start codon	Nucleotide fragment
I.1	3eca-A/ecansba	201	acgc **CAU** UcG **AUG** ucucuaagcugaaugaac
I.2	1cgp-B/eccrp	127	gu **CACUU** cagaga **AAGUG** ggcaaccuggcgu
I.4	1pii/ectrpc	251	cgggugu **UGCU** gggug **AGAA** uaaaguaugug
I.5	1gla-G/ecglyk	254	uugugcgugaaagaagggauggcgaagaaca
I.6	1fia/ecfis	66, 67	gugaugcaa **UACA** cccgu **GGUA** accagaccc
I.8	DHDPS/ecdhdps	225	cauuuugc **CGAG** gca **CGCG** uuauuaaucagc
I.10	1ctf/ecrpobc	55	gacguaauucu **GAAAG** CuG **CUGGC** gcuaaca
I.11	2rsl/ectn1000	146	auagauagagaugcag **UAU** UaA **AUA** uguggc
I.12	1cdd/ecpurmn	100	u **AUGC** CgG **GCGU** uugcugaacauucacccuu
II.1	2abk/ecnth	27, 28	**AGUUCG** ccuuu **UGAAUU** gcugauugccguac
II.2	1dsb/ecdsf	59	uaccacgu **CAAC** uucaug **GGUG** gugaccugg
II.2	1dsb/ecdsf	142, 143, 145	gcagcugacgu **GCAA** uugcg **UGGC** guuccgg
II.3	1dra/ecfolx	28, 29	cucgcc **UGGU** uuaaacgc **AACA** ccuuaaaua
II.4	1aam/ecaspc	324, 325, 326	u **UGUU** cguc **AAUA** cgcugcaggaaaaaggcg
II.6	1trb/ectrxb	117	cgcuaucucggc **CUGC** ccucu **GAAG** aagccu
II.7	1ake/ecadk	126	caugcgccgucuggu **CGUG** uuuau **CACG** uua
II.8	1goa/ecqrnh	66	auuuugaguacc **GACAG** ccagu **AUGUC** cgcc
II.8	1goa/ecqrnh	124, 125	caug **CCGG** acac **CCGG** aaaacgaacgcugug
III.1	2liv/eclivhmg	125	g **GGCC** GaC **GGCU** gccaaauauauucuugaga
III.1	2liv/eclivhmg	334	**UGGC** ac **GCCA** acggcacggccaccgaugcga
III.2	2lbp/eclivhmg	125, 126	g **GGCC** aacggc **GGCA** aaauacauucuugaga
III.2	2lbp/eclivhmg	326, 328, 327	c **UUAA** gggau **UUGA** uuuuggugguuccagu
III.3	1dmb/ecuw89	112	gcguuaucgc **UGAU** uu **AUAA** caaagaucugc
III.3	1dmb/ecuw89	268, 269	gc **CGCCAGU** ccgaacaaaga **GCUGGCG** aaag
III.3	1dmb/ecuw89	317	a **UUGCCG** ccacca **UGGAAA** acgcccagaaag
III.4	2dri/ecrbsp	97, 99, 98	**GUGG** ugag **CCAC** auugcuucugauaacguac
III.4	2dri/ecrbsp	236, 235	cuac **CCGAU** CaG **AUUGG** cgcgaaaggcgucg
III.5	2gbp/ecmglabc	117, 121	auuau **UCAAG** gcga **UUUGA** uugcuaaacacu
III.5	2gbp/ecmglabc	293	guaccu **UAUG** uugg **CGUA** gauaaagacaacc
III.6	1abe/ecarafgh	254	cca **AGCC** cggaccuacau **GGCU** auaaaucca
III.7	1pfk/ecpfka	258	gugccu **UACG** ac **CGUA** uucuggcuucccgua
III.7	1pfk/ecpfka	314	gac **UGC** GcC **GAA** aaaauguau
IV.1	1etu/ectgtufb	197, 198	uc **UUAC** auucc **GGAA** ccagagcgugcgauug
IV.1	1etu/ectgtufb	299, 301	aagccgcacaccaag **UUCG** aauc **UGAA** gugu
IV.2	1kfd/ecpola	323	acggugauuucuua **UGAC** aacuac **GUCA** cca
IV.3	2pol/ecdnaan	100	**CGCU** CcG **GGCG** uagccguuuuucgcugucua
IV.4	1bia/ecbira	66	uuacuuaaug **CUAA** acagaua **UUGG** gucagc
IV.4	1bia/ecbira	270, 271	aauuuuauuaa **UCGC** cca **GUGA** aacuuauca
V.1	3icd/ecicd	123	gaucucuacauc **UGCC** ugcgucc **GGUA** cguu
V.1	3icd/ecicd	159	auuua **UGCG** gguau **CGAA** uggaaagcagacu
V.1	3icd/ecicd	200	ccgug **UUC** GgaA **GAA** ggcaccaaacgucugg
V.2	2glt/ecgshii	134	uuaacg **CCAG** aaacg **CUGG** uuacgcgcaaua
V.2	2glt/ecgshii	202, 198, 197	cca **GCCAU** uaaag **AUGGC** gacaaacgcgugc
V.3	1pda/echemc	107	gccu **UUGU** gucca **AUAA** cuaugacagucugg
V.3	1pda/echemc	196	guaggacaaggugcggug **GGUAU** uga **AUGCC**
VI.1	1rnr/ecnrda	194	uaucucugccugau **GAGCG** uuaa **CGCGC** ugg
VI.1	1rnr/ecnrda	295	ucga **UGAUU** ggucug **AAUAA** agacauucucu
VI.2	2reb/ecreca	44	ucgcu **UUC** AcU **GGA** uaucgcgcuuggggcag
VI.2	2reb/ecreca	333, 332	ucaacgc **CGGA** uuuc **UCUG** uagaugauagcg

[a] Each entry shows the serial number and the PDB/EMBL identifiers as listed in Table 2, followed by the starting codon position of the slow tricodon entries. The nucleotide sequence of the region encompassing 31 nucleotides starting with the base 1 of the tricodon is listed; the subsequences given in bold and uppercase letters can form hairpin stem structures with the intervening subsequence yielding the hairpin loop. Bases indicated by upper case letters in the loop region can also form a base pair.

lowing observations. Translational pauses have been demonstrated experimentally for periplasmic proteins (Randall et al., 1980; Josefsson & Randall, 1981). *E. coli* possesses two targeting pathways either dependent on Sec A/B or on SRP [see (MacFarlane & Muller, 1995) and references therein]. The Sec pathway is thought to translocate proteins in a posttranslational manner (Wickner et al., 1991), whereas the SRP pathway could, by analogy with eukaryotes, possibly function in a cotransla-

Table 8. *Statistical significance of the occurrence of slow regions near cytosolic links versus periplasmic links*

	(len/30) list[a]				(len/40) list				(len/50) list			
	Cytosolic[b]		Periplasmic[b]		Cytosolic		Periplasmic		Cytosolic		Periplasmic	
Space[c] (±)	$o-e$[d] (σ)	l/e[e] %	$o-e$ (σ)	l/e %	$o-e$ (σ)	l/e %	$o-e$ (σ)	l/e %	$o-e$ (σ)	l/e %	$o-e$ (σ)	l/e %
14	1.5	67	1.3	68	2.4	57	2.0	56	2.9	43	1.3	52
12	1.7	62	1.5	64	2.7	52	2.3	56	3.3	40	1.9	52
10	2.6	60	1.5	64	3.1	48	2.1	56	3.5	37	1.8	52
8	2.4	47	1.7	64	2.7	38	2.1	56	2.8	26	1.7	48
6	3.0	42	2.0	56	3.9	36	2.4	48	4.3	24	2.1	40
4	3.4	37	1.6	44	4.2	30	1.7	32	4.6	19	2.0	28
2	3.6	22	2.4	32	4.7	16	2.5	24	4.8	11	2.5	24
0	5.2	11	3.3	16	7.1	7	3.9	12	5.9	6	3.9	12

[a] (*len/n*) list, Number (residue length of protein divided by *n*) of the slowest mRNA regions used in this study for each of the cytosolic or periplasmic proteins indicated.

[b] Cytosolic, data set of 29 cytosolic proteins (Table 2); periplasmic, data set of 8 periplasmic proteins (Table 2).

[c] Space (see footnotes for Table 3 and definition in Fig. 1), the nearness of a slow region to the link/end in terms of the number of separating codons on either side of the link/end (see Materials and methods for its determination).

[d] $o-e$, difference in the observed and expected number of occurrences of slow regions near the links/ends and is given in units of standard deviations (σ) (see Materials and methods).

[e] l/e, percentage of unique links/ends that are delineated by slow regions. Total number of links/ends in cytosolic proteins is 81, and for periplasmic proteins, it is 25.

tional targeting pathway (Phillips & Silhavy, 1992; MacFarlane & Muller, 1995). The release of the protein from the prokaryotic membrane appears to coincide with the folding of the protein (Creighton, 1993). MacFarlane and Muller (1995) further suggest that the role of bacterial SRP is in targeting and subsequent integration of hydrophobic membrane proteins, whereas that of Sec A is in posttranslational targeting of secretory proteins. Thus, for SRP mediated targeting, the folding of proteins into the final structure could still be cotranslational despite being translocated, and hence the proposed pause sites might facilitate the domain folding. Moreover, the prokaryotic nascent chain passes through only a single membrane in contrast to the eukaryotic secretory proteins that are exported to different organelles. For Sec A-mediated targeting, the secretory proteins (that can be translocated after release from the ribosome) are held in a competent state by interaction with Sec B and/or Sec A [both of which are components of the secretory system that includes also Sec Y, Sec E, Sec D, and Sec F (Douville et al., 1995)]. The role of the proposed translational pauses in the secretory proteins may allow interaction of the nascent peptide with Sec A or Sec B, which may be optimal in the presence of a domain. Nonetheless, further investigation regarding the binding sites of Sec A/B is required.

Domain folding

The results presented here suggest that a translational pause occurs on the *E. coli* mRNA after the protein peptide sequence for a domain or a domain-segment is formed on the ribosome; the pause can be enhanced as the synthesized domain terminus passes through the hydrophobic peptide channel on the ribosome. Two exemplary protein structures where mRNA slow regions are associated with domain linking regions are shown in Figure 2. Pauses that occur at the end of a domain may provide a kinetic time scale for the domain to attain its conformation. Folding of nascent proteins on ribosomes have been investigated for both eukaryotic and bacterial proteins. The typical case studies are *E. coli* tryptophan synthase (Fedorov et al., 1992; however, see Tokatlidis et al., 1995 for the limitations of the assay used) and bacteriophage *P22* tail-spike protein (Friguet et al., 1994). The typical eukaryotic proteins demonstrated to fold in an *E. coli* cell-free coupled transcription/translation system are ricin (Kudlicki et al., 1995) and rhodanese (Kudlicki et al., 1994). The *E. coli* DnaK and DnaJ are associated with nascent polypeptide chains of lambda c1857 repressor in translating ribosomes (Gaitanaris et al., 1994). It has also been shown that the bacterial chaperonins GroEL and GroES act on ribosome-bound nascent rhodanese peptide (Tsalkova et al., 1993). The binding of chaperonins to the above-mentioned nascent proteins indicate that their folding process is at least initiated on the ribosome. It has also been suggested that the chaperonins might control the productive folding once a polypeptide domain has been synthesized (Hendrick et al., 1993).

Domains have been often observed to fold independently of the remainder of polypeptide. A few of many examples include lysozyme (Radford et al., 1992), thermolysin (Corbett & Roche, 1986), aspartate transcarbamylase (Maley & Davidson, 1988), *lexA* repressor (Slilaty et al., 1990), exotoxin (Brinkmann et al., 1992), and Tu elongation factor (Nock et al., 1995). Though it is not always clear that single domains (and domain-segments) fold independently, ends of such segments certainly delineate transition from one structural unit to another. However, given the many examples of independent folding, our assumption is reasonable. Furthermore, the correlation between structural domain terminii and slowly translated RNA regions is significant, and thereby supports the presumption. During the growth of the domain on the ribosome, concurrent secondary structural elements may form with some amount of ordering aided by other

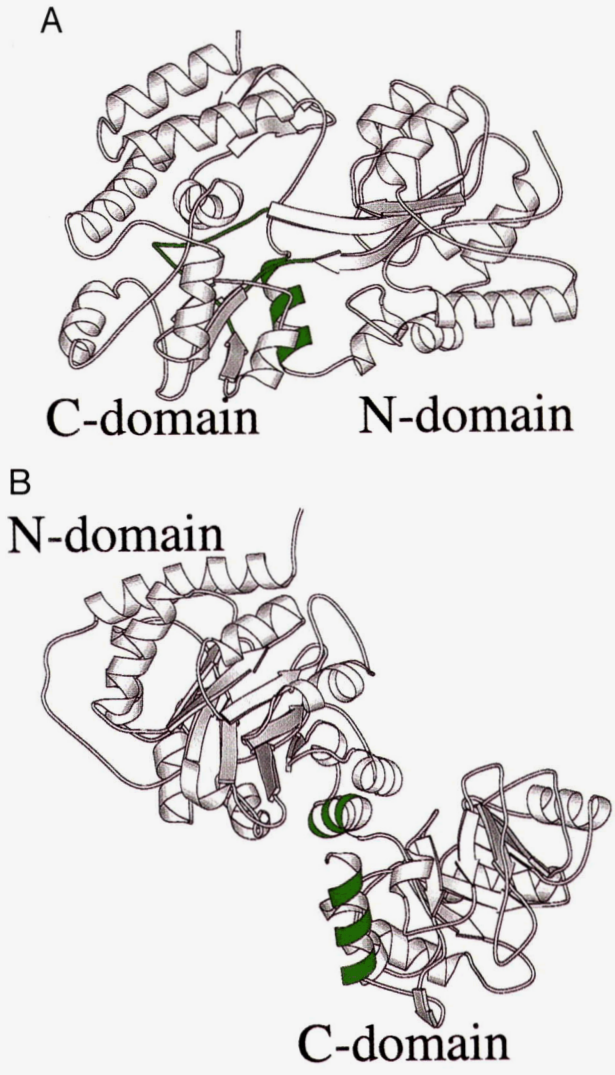

Fig. 2. Mapping of slow regions on the encoded protein's tertiary main-chain fold. Structural elements corresponding to the slow regions are shown in green. **A:** D-Maltodextrin binding protein (1dmb, III.3 in Table 2) with the mapped slow regions corresponding to residues 112–114, 253–255, and 317–319 in the amino acid sequence. **B:** Bifunctional indole-3-glycerol phosphate synthase/N-5′-phospho-ribosyl anthranilate isomerase protein with the mapped slow regions corresponding to residues 251–253 and 442–444. Illustrations showing the path of the backbone topology were generated using MOLSCRIPT (Kraulis, 1991).

slow/fast regions within the domain, and the overall domain organization may well be aided by the pauses suggested here at the domain C-terminus. Translation termination, also a slow process (Lodish & Jacobsen, 1972), may also be important in the establishment of interdomain interactions and thus final domain stabilization.

Mutation data in support of the hypothesis

This analysis would predict that mutagenesis to translationally "fast" codons and/or "non-sticky" amino acids in the slow regions delineating domain ends on mRNA could diminish protein folding, possibly leading to a phenotype displaying temperature sensitivity, some nonfunctional trait, and the like. The protein mutation database [PMD available on the EMBL server via the SRS query system (Etzold & Argos, 1993a, 1993b) with address http://www.embl-heidelberg.de/srs/srsc] was searched regarding the proteins considered in this study, but no synonymous codon substitution mutations were found. Nonetheless, data relating to missense mutations in the *lacI* gene that lead to amino acid replacements in the protein were available. The results of the analysis are given subsequently.

The len/30 list of slowest nucleotide codon regions for the complete lacI repressor protein is as follows: (1) 188–191 (PLSS); (2) 332–335 (PNTQ); (3) 319–322 (LPVS); (4) 279–282 (SSCY); (5) 112–115 (HNLL); (6) 119–121 (VSG); (7) 146–148 (LFL); (8) 220–222 (WSA); (9) 77–79 (SQI); (10) 226–228 (FQQ); (11) 208–210 (NQI); and (12) 344–347 (DSLM). The structure of the complete repressor has been published recently (Lewis et al., 1996; Matthews, 1996). The four domains are the DNA-binding domain (1–45), the N-terminal subdomain of the core (63–161 and 293–318), the C-terminal subdomain of the core (164–290 and 322–324), and the tetramerization domain (325–360). It is to be noted that the range of domain-segments for the core domains differ in some detail from the model of Matthews and coworkers (Nichols et al., 1993) as given in Table 2 (II.5 1ltp/eclaci). For the discussion here, we consider the latest definitions of the amino acid ranges forming the domains; accordingly, the links/ends become 46–62, 162–164, 291–293, 319–321, 325–327, and 357–360. Because the domain-segment 322–324 is short, two of these links (319–321 and 325–327) can be replaced by one link (325–327). Hence, there are four links and one end (46–62, 162–164, 291–293, 325–327, and 357–360). The slow regions from the len/30 list that are associated with these links/ends are 77–79 separated from the link 46–62 by +14 codons, 146–148 separated from 162–164 by −13 codons, 279–282 separated from 291–293 by −8 codons, 319–322 separated from 325–327 by −2 codons, and 344–347 separated from the end by −9 codons. Missense mutational sites in the *lacI* gene leading to the I^- phenotype (Gordon et al., 1988) occurred in the slow regions associated with all the links; namely, positions 77–78 from 77–79, SQI (S to T or L; Q to E); 148 from 146–148, LFL (L to P); all four residues of 279–282, SSCY (to RPRD).

Elongation factor EF-Tu (1etu/ectgtufb, IV.1 in Table 2) is a three-domain protein with links 199–201 and 298–300 and the C-terminal domain end at 391–393. The len/30 list indicated the corresponding slow codon regions to be 196–200 (for D1) and 299–303 (for D2). Examination of the reported single amino acid substituted mutations (Hwang et al., 1992) revealed three mutations, namely, at positions 196 (D to G), 197 (S to F), and 199 (I to V), that are associated with the slow region for the link 199–201. These suppression mutations significantly affect EF-Tu's ability to interact with EF-Ts. The simulation inversion protein factor fis (I.6 1fia/ecfis in Table 2) is a two-domain protein with link at 70–73 and the C-terminal domain end at 96–98. The link is represented by a slow codon region 63–70. Diethylsulfate-induced missense mutation at codon position 70 leading to an amino acid replacement from Thr to Ile, leads to considerable reduction in the stimulation of Hin-mediated DNA inversion both in vivo and in vitro as well as in the stimulation of lambda excision and binding to enhancer DNA and to the lambda Fis site. Further the mutation obtained from *mutD5 E. coli* strain at the codon location 63, leading to a Leu to Ser substitution, has resulted in a dramatic reduction in all the above-mentioned

functional abilities, and it has been suggested that this mutant has a significantly altered structure (Osuna et al., 1991).

This study has revealed that the boundaries of domains or domain-segments in the protein tertiary structure are coded by likely slowly translated regions on mRNA and are preferentially composed of hydrophobic amino acids that may "stick" to the ribosome tunnel. The codon usage in such slow regions is in agreement with the reported experimental data on the translation rates of codons. The slow regions on mRNA show the potential to form higher-order nucleotide structures. Because experimental data on protein folding and translation rates at specific mRNA regions are not yet complete, the correlation found here is all the more remarkable. RNA is thus likely to not only possess information regarding the primary structure of the protein, but also to select for codons to aid in protein folding and domain (and subdomain) organization at the ribosome. Brunak et al. (1994) identified strong signals in mRNA sequence regions preceding helices and sheets on the encoded proteins and thereby reported a correlation between protein secondary structure and the mRNA nucleotide sequence. Lobry and Gautier (1994) observed that the translational constraints that are known to affect the "genotype" of proteins, are sufficient to affect their "phenotype." The results reported in the present study support further the notion that nucleotide sequences carry numerous superimposed messages (Trifonov, 1989). The results point to mutation experiments that can test the relationship between domain boundaries and translation rate. The specificity of the protein sequence and structural regions to alter as well as the codons to select in the corresponding mRNA segments has not heretofore been so pinpointed. The preferred amino acids and nucleotide codons in the domain linker regions specified in this work should prove useful in the design and engineering of proteins for both in vivo and in vitro systems, as well as in the prediction of domain and subdomain linker regions.

Materials and methods

Codon fractional frequency values

The frequencies of the 61 sense codons for the 20 amino acids are calculated as the fractions of the sum of all codons in mRNAs of highly expressed genes (Sharp & Li, 1987) (namely, *lpp*, *ompA*, *ompC*, *ompF*, *tufA*, *tufB*, *tsf*, *fusA*, *recA*, *dnaK*, and those of ribosomal proteins) from *E. coli*. The nucleotide sequences were extracted from the EMBL nucleotide sequence data bank (Rice et al., 1993) with the aid of the SRS information retrieval computer system (Etzold & Argos, 1993a, 1993b). Only the internal coding regions of the mRNAs (codons 21 through 3 codons prior to the stop codon) were considered because translation-initiation "enhancer" signals occur anywhere in the first 17 codons (McCarthy & Gualerzi, 1994); the minor codons are preferentially located in the 5′ coding region (Chen & Inouye, 1990); and the signal for translation-termination extends to two codon positions 5′ to the stop codon (Brown et al., 1990). Sequences of length less than 50 codons, sequences with conflicts in nucleotide determination, and partially determined nucleotide sequences were not considered. The final size of the database used to calculate the above codon frequency values is 9,512 codons from 60 mRNA sequences corresponding to the above-mentioned highly expressed genes. Thus, any value in Table 1 multiplied by 9,512 yields the actual number of times that a particular codon appears in the coding regions considered for these 60 mRNAs.

Data on protein domains and links/ends

The data set of 37 *E. coli* multidomain proteins with known tertiary structures consisted of 88 such domains comprised by 112 domain-segments (Table 2). The span of amino acid residues between the end of a domain-segment and the beginning of the succeeding domain-segment forms the connecting link; in case such a region falls short of 3 residues, it was constituted by extra residues from the succeeding domain-segment to achieve a length of 3. The C-terminal 3 residues formed the ends of C-terminal domain-segments, i.e., domain ends.

CAI

The CAI, as formulated by Sharp and Li (1987), is an indicative measure of the level of expression of a gene. The index relies on a reference set of highly expressed genes to assess the relative merit of each codon. A score for a given gene is calculated from the frequency of codon usage in the gene compared with that in the reference set. The index thus assesses the extent to which natural selection for optimal (or abundant) codons has acted on the given gene. A high (low) value for the index indicates that a gene is expressed at a high (low) level. The values were determined by using the formalism of Sharp and Li (1987) and the codon frequency values calculated here.

Search for slow regions on mRNA

For the reasons mentioned earlier (in "Rationale"), it was decided to construct a list containing len/30 (residue length of protein divided by 30) number of slowest nucleotide regions for each of the 37 mRNA sequences associated with the known protein structures. A tricodon could be considered as the minimum length of such regions because three successive residues on average constitute basic protein structural motifs; namely, a small beta strand at 3 residues, one helical turn at 3.6, a reverse beta turn at 3–4 amino acids (Colloc'h et al., 1993). In each of the mRNA sequences, the slowness of all individual tricodons starting at each codon position of the mRNA was scrutinized by taking the sum of the codon fractional frequency values (as listed in Table 1) of the constituent codons. The tricodons were arranged in the ascending order of the sum of the frequency values (as listed in Table 1) of their constituent codons. Such an arrangement shows the tricodons in the decreasing order of slowness for translation. For reasons mentioned previously, the tricodon entries falling within the first 20 codons on the mRNA were ignored. The ordered tricodon list was subsequently used to construct the len/30 list. Each tricodon sequentially selected from the above arrangement formed a distinct region in the len/30 list, where tricodons that differed in location by 1–3 codons were strung together to form a single distinct region (e.g., tricodons starting at locations 15 and 18 or 15 and 17 can be combined because the codons 15 to 20 or 15 to 19 form a contiguous region on the mRNA).

Scanning the len/30 list of slow regions for regions near to the domain links/ends

The len/30 list was scanned (Fig. 1) for regions that occurred near mRNA regions corresponding to the protein domain links and ends. The nearness of a slow region with reference to links/ends is quantified by the parameter "*space*," which gives the number of codons that separate the link/end from the slowest tricodon. For reasons of statistical calculations, it also includes the tricodon length. For the distinct region that possesses more than one tricodon, position of the slowest codon within it is used to calculate the value for *space*. The value is calculated as per the following formula:

If the termini of a link or an end are denoted by L1, L2, and those of the slowest tricodon in an entry of the len/30 list by s1, s2, then *space* is (s1 − L1) when s1 < L1; or *space* is 0 when s1 = L1 (because the minimum length of link/end is 3) or when s1 > L1 and s2 ≤ L2; or *space* is (s2 − L2) when s1 > L1 and s2 > L2.

For each of the 37 proteins, the above exercise was carried out (as illustrated in Fig. 1) to determine the number of len/30 slowest regions (*ndmr*) that occur on mRNA near or at the regions corresponding to protein domain links/ends by a maximum space of ±16 codons and the number of slowest regions (*nre*) required to be considered from the len/30 list to observe the *ndmr* regions. The number of unique links/ends (*l/e*) represented by the *ndmr* regions was also elicited. These three quantities were summed up individually over the 37 proteins to yield overall *ndmr*, *nre*, and *l/e* values. Each of the considered proteins that had at least one of its links/ends represented in the len/30 list was enumerated by the parameter *np*. All the abovementioned parameters (*nre*, *ndmr*, *l/e*, and *np*) were determined for different *space* values and also for various values of *n* in (*len/n*) lists.

Statistical significance of occurrence of slow regions near links/ends

As listed in Table 2, the number of proteins considered is 37 and the number of domain-segments is 112, of which 75 are either N-terminal or internal fragments, whereas 37 are C-terminal protein segments. The length of six C-terminal domain-segments (Table 2) does not exceed 15 residues. Because pausing/stacking of the ribosome on mRNA is likely to occur at translation-termination (Wolin & Walter, 1988), which is the slowest process (Lodish & Jacobsen, 1972) in translation, and because the *E. coli* ribosome protects 10–15 codons on mRNA (Robertson et al., 1973; Gold et al., 1981), these latter six cases were not considered here; the translation-termination process could well act as a pause for the folding of such small domain-segments. Consequently, the effective number of links/ends becomes 106. The total residue lengths of these links/ends and the 37 proteins are 424 and 11,225, respectively. Because slow regions are not considered in the first 20 codons of each mRNA, the effective total protein residue length is 11,225 − (37 * 20) = 10,485. The length of the search regions for a given "*space*" is 424 + [75 * (2 * *space*)] + [31 * *space*] where * indicates multiplication. The number 424 corresponds to the total residue length of links/ends; 75 is the number of interdomain links; 31, the number of C-terminal ends; and the term [2 * *space*] is used because the search for slow regions is performed at both sides of links, whereas for ends only one side is possible.

The expected frequency (*expfreq*) for the occurrence of slow regions with a given "*space*" is the length of the search region divided by the effective total residue length of all proteins in the data set. The statistical significance ($o - e$) of observing *ndmr* slowest regions from *nre* slowest regions taken from a list with *len/n* members is expressed by the difference between the observed and expected numbers in terms of σ. This statistical parameter ($o - e$) is calculated from the following: *expected ndmr* = *nre* * *expfreq*; $\sigma = [nre * expfreq * (1 - expfreq)]^{1/2}$; $o - e = (ndmr - \text{expected } ndmr)/\sigma$. The meaning of σ and ($o - e$) can be amplified. The term *expfreq* indicates the probability of selecting a position from search regions in a random experiment. When the random experiment is repeated *nre* times, *nre* * *expfreq* hits would be expected in the search regions. The σ in the expected number over *nre* number of trials is the square root of *nre* * *expfreq* * (1 − *expfreq*). The term ($o - e$) relates the over- or underrepresentation in the actual occurrences related to the expected occurrences in terms of σ such that ($o - e$) with an absolute value of 2 or more is considered significant.

Solvent accessibility of residues in the slow regions

For the calculation of solvent accessibility of the residues forming slow regions, the accessibility parameter as listed in the DSSP secondary structure assignment files (Kabsch & Sander, 1983) was taken. This parameter is the solvent-accessible surface area of the individual residues expressed in Å^2. It is that area on a protein's van der Waals atomic surface to which a water molecule of radius 1.4 Å would have access. The solvent-accessible surface itself is delineated by the area available to the solvent probe center. Such calculations are possible only when the three-dimensional structure of the protein is known.

Synonymous codon usage and relative synonymous codon usage

The synonymous codon frequency or usage of a particular codon for a given amino acid type is defined as frequency of the occurrence of that codon over the total number of the codons observed for the given coded amino acid in a protein data set. Relative synonymous codon usage (*RSCU*), which indicates the relative use of a codon over its synonymous codons, was calculated as defined by Sharp and Li (1986, 1987); namely, *RSCU* = (synonymous codon frequency)/(expected frequency of the codon), where the expected frequency of a codon = (1/*n*) and *n* is the number of synonymous codons that code for the given amino acid. We designated codons as *rare* when their *RSCU* values as calculated from the mRNAs of highly expressed genes used in Table 1 was ≤ 0.08. The next highest values were 0.13 and 0.19. Sharp and Li (1986, 1987) considered rare codons to be those with *RSCU* < 0.05 (the next highest values in their data set were 0.10 and 0.11).

Comparison of synonymous codon usage in slow regions and in highly expressed genes

The synonymous codon usage of each of the 61 sense codons in highly expressed genes is designated as *sr* and that in slow regions as *sd* and in search regions (region encompassing the mRNA spans corresponding to links/ends along with 12 codons on either side; for ends only upstream region considered) as *sl*.

To illustrate the behavior of codon usage in slow regions compared with that in highly expressed genes or in search regions and also to assess the avoidance or preference of a synonymous codon in slow regions or in highly expressed genes, three parameters were defined: *srx* for highly expressed genes, *sdx* for slow regions, and *slx* for search regions. They are calculated as follows: $srx = sr - X$, $sdx = sd - X$, and $slx = sl - X$, where X is the expected frequency of a synonymous codon assuming that all synonymous codons for an amino acid are employed with equal probability such that $X = (1/n)$ where n is the number of synonymous codons that code for the same amino acid. A value of zero for *srx*, *sdx*, or *slx* points to neutral usage, a negative value indicates avoidance of that codon, and a positive value shows preference. If the *srx* and *sdx* values for a codon are different in sign, then a reversal in the trend of preference/avoidance of the codon has taken place in slow mRNA regions relative to its usage observed in highly expressed genes.

Translation rates of individual codons

Curran and Yarus (1989) have reported the rate of translation for 29 sense codons. Bonekamp et al. (1989) have also given such values for 12 codons, of which 8 are also reported in Curran and Yarus (1989). For the purpose of comparing the codon usage in slow regions with the reported translation rates, we used the values given by Curan and Yarus (1989). Their measurements yield relative rates of association between the codon and the cognate aminoacyl-tRNA; nonetheless, extrapolations to the normal translocation step (or the relative movement of ribosome on mRNA) are reasonable. However, selection of aminoacyl-tRNA may be the rate-limiting step as revealed by the differences in the expression of genes containing rare codons in rich versus poor glucose-containing medium (Del Tito et al., 1995).

Search for RNA hairpin stem structures

Along with the regular WC base pairs (U-A, A-U, C-G, G-C), other common noncanonical base pairs (U.G, G.U, A.G, and G.A) were also allowed but only if a single base pair was noncanonical in a stem of minimal length four base pairs. In ribosomal RNA molecules, the most prevalent hairpin loop is a tetra-loop, with the next most observed as a penta-loop (Woese et al., 1990). Evolutionary substitution of tetra-loops by tri- and penta-loops have been observed. Thus, we allowed the length of hairpin loops to be 3–6 nucleotides. In the absence of such hairpin stems involving nucleotides from the slow region, the immediate downstream 5–9 nucleotides were also searched for hairpin stems in accordance with the observations of secondary structure starting downstream of the frameshift site on the *pol* mRNA (Tu et al., 1992).

Acknowledgments

T.A.T. thanks the International Human Frontier Science Program Organization for a long-term fellowship (LT-82/94). We thank the editor and the referees for useful suggestions and comments.

References

Andersson SGE, Kurland CG. 1990. Codon preferences in free-living microorganisms. *Microbiol Rev 54*:198–210.

Argos P. 1990. An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. *J Mol Biol 211*:943–958.

Bairoch A, Apweiler R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res 24*:21–25.

Baldwin RL. 1975. Intermediates in protein folding reactions and the mechanism of protein folding. *Annu Rev Biochem 44*:453–475.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. *J Biol Chem 257*: 3026–3031.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bonekamp F, Andersen HD, Christensen T, Jensen KF. 1985. Codon-defined ribosomal pausing in *Escherichia coli* detected by using the *pyrE* attenuator to probe the coupling between transcription and translation. *Nucleic Acids Res 13*:4113–4123.

Bonekamp F, Dalboge H, Christensen T. 1989. Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilisation in *Escherichia coli*. *J Bacteriol 171*:5812–5816.

Bordo D, Argos P. 1994. The role of side-chain hydrogen bonds in the formation and stabilisation of secondary structure in soluble proteins. *J Mol Biol 243*:504–519.

Bowie JU, Clarke ND, Pabo CO, Sauer RT. 1990. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins Struct Funct Genet 7*:257–264.

Bowie JU, Sauer RT. 1989. Identifying determinants of folding and activity for a protein of unknown structure. *Proc Natl Acad Sci USA 86*:2152–2156.

Brimacombe R. 1995. The structure of ribosomal RNA: A three-dimensional jigsaw puzzle. *Eur J Biochem* 230:365–383.

Brinkmann U, Buchner J, Pastan I. 1992. Independent domain folding of *Pseudonomas* exotoxin and single-chain immunotoxin: Influence of interdomain connections. *Proc Natl Acad Sci USA 89*:3075–3079.

Brown CM, Stockwell PA, Trotman CNA, Tate WP. 1990. The signal for the termination of protein synthesis in procaryotes. *Nucleic Acids Res 18*:2079–2086.

Brunak S, Engelbrecht J, Kesmir C. 1994. Correlation between protein secondary structure and the mRNA nucleotide sequence. In: Bohr H, Brunak S, eds. *Protein structure by distance analysis*. Amsterdam, IOS Press. pp 327–334.

Carter PW, Bartkus JM, Calvo JM. 1986. Transcription attenuation in *Salmonella typhimurium*: The significance of rare leucine codons in the leu leader. *Proc Natl Acad Sci USA 83*:8127–8131.

Chaney WG, Morris AG. 1979. Non-uniform size distribution of nascent peptides—The effect of messenger RNA structure upon the rate of translation. *Arch Biochem Biophys 194*:283–291.

Chavancy G, Chevallier A, Fournier A, Garel JP. 1979. Adaptation of isotRNA concentration to mRNA codon frequency in the eukaryotic cell. *Biochimie 61*:71–78.

Chen GT, Inouye M. 1990. Suppression of the negative effect of minor arginine codons on gene expression; preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. *Nucleic Acids Res 18*:1465–1473.

Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP. 1993. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. *Protein Eng 6*:377–382.

Corbett RJ, Roche RS. 1986. Independent folding of autolytic fragments of thermolysin and their domain-like properties. *Int J Pept Protein Res 28*:549–559.

Creighton TE. 1993. Biosynthesis of proteins. In: *Proteins—Structures and molecular properties*. New York, WH Freeman and Company. pp 49–104.

Curran JF, Yarus M. 1989. Rates of aa-tRNA selection at 29 sense codons in vivo. *J Mol Biol 209*:65–67.

Del Tito BJ Jr, Ward JM, Hodgson J, Gershater CJL, Edwards H, Wysocki LA, Watson FA, Sathe G, Kane JF. 1995. Effects of a minor isoleucyl tRNA on heterologous protein translation in *Escherichia coli*. *J Bacteriol 177*:7086–7091.

Dorit RL, Schoenbach L, Gilbert W. 1990. How big is the universe of exons? *Science 250*:1377–1382.

Douville K, Price A, Eichler J, Economou A, Wickner W. 1995. Sec YEG and Sec A are the stoichiometric components of preprotein translocase. *J Biol Chem 270*:20106–20111.

Edelman GM, Cunningham BA, Gall We, Gottlieb PD, Rutishauser U, Waxdal MJ. 1969. The covalent structure of an entire gammaG immunoglobulin molecule. *Proc Natl Acad Sci USA 63*:78–85.

Etzold T, Argos P. 1993a. Transforming a set of biological flat file libraries. *CABIOS 9*:59–64.

Etzold T, Argos P. 1993b. SRS – An indexing and retrieval tool for flat file data libraries. *CABIOS 9*:49–57.
Farabaugh PJ. 1996. Programmed translational frameshifting. *Microbiol Rev 60*:103–134.
Fedorov AN, Friguet B, Djavadi-Ohaniance L, Alakhov YB, Goldberg ME. 1992. Folding on the ribosome of *Escherichia coli* tryptophan synthase beta subunit nascent chains probed with a conformation-dependent monoclonal antibody. *J Mol Biol 228*:351–358.
Ficket JW. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res 10*:5303–5318.
Friguet B, Djavadi-Ohaniance L, King J, Goldberg ME. 1994. In vitro and ribosome-bound folding intermediates of *P22* tailspike protein detected with monoclonal antibodies. *J Biol Chem 269*:15945–15949.
Gaitanaris GA, Vysokanov A, Hung SC, Gottesman ME, Gragerov A. 1994. Successive action of *Escherichia coli* chaperones in vivo. *Mol Microbiol 14*:861–869.
Gesteland RF, Weiss RB, Atkins JF. 1992. Recoding: Reprogrammed genetic decoding. *Science 257*:1640–1641.
Gold L, Pribnow D, Schneider T, Shinedling S, Singer BS, Stormo G. 1981. Translation-initiation in prokaryotes. *Annu Rev Microbiol 35*:365–403.
Goldman E, Rosenberg AH, Zubay G, Studier FW. 1995. Consecutive low-usage leucine codons block translation only when near the 5′ end of a message in *Escherichia coli*. *J Mol Biol 245*:467–473.
Gordon AJE, Burns PA, Fix DF, Yatagai F, Allen FL, Horsfall MJ, Halliday JA, Gray J, Bernelot-Moens C, Glickman BW. 1988. Missense mutation in the *lacI* gene of *Escherichia coli* – Inferences on the structure of the repressor protein. *J Mol Biol 200*:239–251.
Gouy M, Gautier C. 1982. Codon usage in bacteria: Correlation with gene expressivity. *Nucleic Acids Res 10*:7055–7074.
Gribskov M, Devereux J, Burgess RR. 1984. The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res 12*:539–549.
Grosjean H, Fiers W. 1982. Preferential codon usage in prokaryotic genes: The optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene 18*:199–209.
Gutman GA, Hatfield GW. 1989. Nonrandom utilization of codon pairs in *Escherichia coli*. *Proc Natl Acad Sci USA 86*:3699–3703.
Hardesty B, Odom OW, Picking W. 1992. Ribosome function determined by fluorescence. *Biochimie 74*:391–401.
Harms E, Umbarger HE. 1987. Role of codon choice in the leader region of the *ilvGMEDA* operon of *Serratia marcescens*. *J Bacteriol 169*:5668–5677.
Hendrick JP, Langer T, Davis TA, Hartl FV, Wiedmann M. 1993. Control of folding and membrane translocation by binding of the chaperone DnaJ to nascent polypeptides. *Proc Natl Acad Sci USA 90*:10216–10220.
Hwang YW, Carter M, Miller DL. 1992. The identification of a domain in *Escherichia coli* elongation factor Tu that interacts with elongation factor Ts. *J Biol Chem 267*:22198–22205.
Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J Mol Biol 151*:389–409.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol 2*:13–34.
Ikemura T, Ozeki H. 1983. Codon usage and transfer RNA contents: Organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harbor Symp Quant Biol 47*:1087–1097.
Josefsson LL, Randall LL. 1981. Different exported proteins in *E. coli* show differences in the temporal mode of processing in vivo. *Cell 25*:151–157.
Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.
Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amin acids. *Science 262*:1680–1685.
Kane JF, Violand BN, Curran DF, Staten NR, Duffin KL, Bogosian G. 1992. Novel in-frame two codon translational hop during synthesis of bovine placental lactogen in a recombinant strain of *Escherichia coli*. *Nucleic Acids Res 20*:6707–6712.
Kim J, Klein PG, Mullet JE. 1991. Ribosomes pause at specific sites during synthesis of membrane-bound chloroplast reaction center protein D1. *J Biol Chem 266*:14931–14938.
Kim JK, Hollingsworth MJ. 1992. Localization of in vivo ribosome pause sites. *Anal Biochem 206*:183–188.
Kinnaird JH, Burns PA, Fincham JR. 1991. An apparent rare-codon effect on the rate of translation of a *Neurospora* gene. *J Mol Biol 221*:733–736.
Krasheninnikov IA, Komar AA, Adzhubei IA. 1989. Role of the code redundancy in determining cotranslational protein folding. *Biokhimiia 54*:187–200.
Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.
Kudlicki W, Kitaoka Y, Odom OW, Kramer G, Hardesty B. 1995. Elongation and folding of nascent ricin chains as peptidyl-tRNA on ribosomes: The effect of amino acid deletions on these processes. *J Mol Biol 252*:203–212.
Kudlicki W, Odom OW, Kramer G, Hardesty B. 1994. Chaperone-dependent folding and activation of ribosome-bound nascent rhodanese. *J Mol Biol 244*:319–331.
Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brenner RG, Lu P. 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science 271*:1247–1254.
Lipman DJ, Wilbur WJ. 1983. Contextual constraints on synonymous codon choice. *J Mol Biol 163*:363–376.
Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res 22*:3174–3180.
Lodish HF, Jacobsen M. 1972. Regulation of hemoglobin synthesis – Equal rates of translation and termination of α- and β-globin chains. *J Biol Chem 247*:3622–3629.
MacFarlane J, Muller M. 1995. The functional integration of a polytopic membrane protein of *Escherichia coli* is dependent on the bacterial signal-recognition particle. *Eur J Biochem 233*:766–771.
Maley JA, Davidson JN. 1988. The aspartate transcarbamylase domain of a mammalian multifunctional protein expressed as an independent enzyme in *Escherichia coli*. *Mol Gen Genet 213*:278–284.
Matthews KS. 1996. The whole lactose repressor. *Science 271*:1245–1246.
McCarthy JEG, Gualerzi C. 1994. Prokaryotic translation: The interactive pathway leading to initiation. *Trends Genet 10*:402–407.
McNally T, Purvis IJ, Fothergill-Gilmore LA, Brown AJP. 1989. The yeast pyruvate kinase gene does not contain a string of non-preferred codons: Revised nucleotide sequence. *FEBS Lett 247*:312–316.
Mirwaldt C, Korndorfer I, Huber R. 1995. The crystal structure of dihydrodipicolinate synthase from *Escherichia coli* at 2.5 Å resolution. *J Mol Biol 246*:227–239.
Nichols JC, Vyas NK, Quiocho FA, Matthews KS. 1993. Model of lactose repressor core based on alignment with sugar-binding proteins is concordant with genetic and chemical data. *J Biol Chem 268*:17602–17612.
Nock S, Grillenbeck N, Ahmadian MR, Ribeiro S, Kreutzer R, Sprinzl M. 1995. Properties of isolated domains of the elongation factor Tu from *Thermus thermophilus* HB8. *Eur J Biochem 234*:132–139.
Osuna R, Finkel SE, Johnson RC. 1991. Identification of two functional regions in Fis: The N-terminus is required to promote Hin-mediated DNA inversion but not lambda excision. *EMBO J 10*:1593–1603.
Phillips DC. 1966. The three-dimensional structure of an enzyme molecule. *Sci Am 215*:78–90.
Phillips GJ, Silhavy TJ. 1992. The *E. coli ffh* gene is necessary for viability and efficient protein export. *Nature 359*:744–746.
Picking WD, Odom OW, Tsalkova T, Serdyuk I, Hardesty B. 1991. The conformation of nascent polylysine and polyphenylalanine peptides on ribosomes. *J Biol Chem 266*:1534–1542.
Post LE, Nomura M. 1980. DNA sequences from the *str* operon of *Escherichia coli*. *J Biol Chem 255*:4660–4666.
Purvis IJ, Bettany AJE, Santiago TC, Coggins JR, Duncan K, Eason R, Brown AJP. 1987. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. *J Mol Biol 193*:413–417.
Radford SE, Dobson CM, Evan PA. 1992. The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature (Lond) 358*:302–307.
Randall LL, Josefsson LG, Hardy SJ. 1980. Novel intermediates in the synthesis of maltose binding protein in *Escherichia coli*. *Eur J Biochem 107*:375–379.
Rice CM, Fuchs R, Higgins DG, Stoehr PJ, Cameron GN. 1993. The EMBL data library. *Nucleic Acids Res 21*:2967–2971.
Robertson HD, Barrell BG, Weith HL, Donelson JE. 1973. Isolation and sequence analysis of a ribosome-protected fragment from bacteriophage ΦX174 DNA. *Nature New Biol 241*:38–40.
Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, Millican A, Eaton M, Humphreys E. 1984. Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Res 12*:6663–6671.
Rosenberg AH, Goldman E, Dunn JJ, Studier FW, Zubay G. 1993. Effects of consecutive AGG codons on translation in *Escherichia coli* demonstrated with a versatile codon test system. *J Bacteriol 175*:716–722.
Rossmann MG, Moras D, Olsen KW. 1974. Chemical and biological evolution of nucleotide-binding protein. *Nature 250*:194–199.
Sharp PM, Li WH. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for rare codons. *Nucleic Acids Res 14*:7737–7749.

Sharp PM, Li W. 1987. The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res 15*:1281–1295.

Sharp PM, Tuohy TMF, Mosurski KR. 1986. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res 14*:5125–5143.

Shpaer EG. 1985. The secondary structure of mRNAs from *Escherichia coli*: Its possible role in increasing the accuracy of translation. *Nucleic Acids Res 13*:275–288.

Shpaer EG. 1986. Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. *J Mol Biol 188*:555–564.

Siemion IZ, Siemion PJ. 1994. The informational context of the third base in amino acid codons. *Biosystems 33*:139–148.

Slilaty SN, Ouellet M, Fung M, Shen SH. 1990. Independent folding of individual components in hybrid proteins. Evidence that the carboxy terminal 135 residues of the *Lex A* repressor constitute a single autonomous domain. *Eur J Biochem 194*:103–108.

Sorensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J Mol Biol 207*:365–377.

Sorensen MA, Pedersen S. 1991. Absolute in vivo translation rates of individual codons in *Escherichia coli*—The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J Mol Biol 222*:265–280.

Spanjaard RA, van Duin J. 1988. Translation of the sequence AGG-AGG yields 50% ribosomal frameshift. *Proc Natl Acad Sci USA 85*:7967–7971.

Spanjaard RA, Chen K, Walker JR, van Duin J. 1990. Frameshift suppression at tandem AGA and AGG codons by cloned tRNA genes: Assigning a codon to *argU* tRNA and T4 tRNA Arg. *Nucleic Acids Res 18*: 5031–5036.

Staden R, McLachlan AD. 1982. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res 10*:141–156.

Taylor FJR, Coates D. 1989. The code within codons. *Biosystems 22*:177–187.

Tokatlidis K, Friguet B, Deville-Bonne D, Baleux F, Fedorov AN, Navon A, Djavadi-Ohaniance L, Goldberg ME. 1995. Nascent chains: Folding and chaperone interaction during elongation on ribosome. *Phil Trans R Soc Lond B Biol Sci 348*:89–95.

Trifonov EN. 1989. The multiple codes of nucleotide sequences. *Bull Math Biol 51*:417–432.

Tsalkova T, Zardeneta G, Kudlicki W, Kramer G, Horowitz PM, Hardesty B. 1993. GroEL and GroES increase the specific enzymatic activity of newly-synthesised rhodanese if present during in vitro transcription/translation. *Biochemistry 32*:3377–3380.

Tu C, Tzeng TH, Bruenn JA. 1992. Ribosomal movement impeded at a pseudoknot required for frameshifting. *Proc Natl Acad Sci USA 89*: 8636–8640.

van-den-Broek LA, Lazaro E, Zylicz Z, Fennis PJ, Missler FA, Lelieveld P, Garzotto M, Wagener DJ, Ballesta JP, Ottenheijm HC. 1989. Lipophilic analogues of sparsomycin as strong inhibitors of protein synthesis and tumor growth: A structure–activity relationship study. *J Med Chem 32*: 2002–2015.

Varenne S, Baty D, Verheij H, Shire D, Lazdunski C. 1989. The maximum rate of gene expression is dependent on the downstream context of unfavourable codons. *Biochimie 71*:1221–1229.

Varenne S, Buc J, Lloubes R, Lazdunski C. 1984. Translation is a non-uniform process—Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol 180*:549–576.

Volkenstein MV. 1966. The genetic coding of the protein structure. *Biochim Biophys Acta 119*:421–424.

Wickner W, Driessen AJM, Hartl FU. 1991. The enzymology of protein translocation across the *Escherichia coli* plasma membrane. *Annu Rev Biochem 60*:101–124.

Wiedmann B, Sakai H, Davis TA, Wiedmann M. 1994. A protein complex required for signal-sequence-specific sorting and translocation. *Nature 370*:434–440.

Woese CR, Dugre DH, Saxinger WC, Dugre SA. 1966. The molecular basis of the genetic code. *Proc Natl Acad Sci USA 55*:966–974.

Woese CR, Winker S, Gutell RR. 1990. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops." *Proc Natl Acad Sci USA 87*:8467–8471.

Wolin SL, Walter P. 1988. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J 7*:3559–3569.

Yamao F, Andachi Y, Muto A, Ikemura T, Osawa S. 1991. Levels of tRNAs in bacterial cells as affected by aminoacid usage in proteins. *Nucleic Acids Res 19*:6119–6122.

Yarus M, Folley LS. 1985. Sense codons are found in specific contexts. *J Mol Biol 182*:529–540.

Yonath A. 1992. Approaching atomic resolution in crystallography of ribosomes. *Annu Rev Biophys Biomol Struct 21*:77–93.

Yue K, Dill KA. 1992. Inverse protein folding problem: Designing polymer sequences. *Proc Natl Acad Sci USA 89*:4163–4167.

Zhang S, Goldman E, Zubay G. 1994. Clustering of low usage codons and ribosome movement. *J Theor Biol 170*:339–354.